

Deciphering regulation in *Escherichia coli*: from genes to genomes

Thesis by
Suzannah Michelle Beeler

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy in Biology



CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2022
Defended July 22, 2021

© 2022

Suzannah Michelle Beeler
ORCID: 0000-0002-1930-4827

Some rights reserved. This thesis is distributed under a Creative Commons
Attribution License CC-BY 4.0.

ACKNOWLEDGEMENTS

First and foremost I must thank my PhD advisor, Rob Phillips. I have learned so much from him during the time I've spent with him in the lab and in the classroom. He has instilled in me the immense value of biological numeracy and the ability to tackle problems in biology from pure thought alone. I can safely say that my PhD was far from typical, but in a way that has prepared me for my next steps better than I could have ever imagined. With Rob I have been given the opportunity to teach at the Marine Biological Laboratory, at GIST in Korea, and even in New Zealand and the Galápagos with the Caltech Alumni. With these innumerable teaching experiences, I eagerly await the next stage of my academic career where I intend to spread the immeasurable value of a quantitative understanding of biology to my future students.

Within the realm of teaching quantitative biology, Justin Bois has also been an immense source of inspiration. Within my first years at Caltech, I had the pleasure of taking four classes from him, imbuing me with skills that I used throughout the remainder of my graduate career. And while I only had the opportunity to TA with him once, I will remain inspired by his thoughtful and rigorous approach to teaching. When explaining my future job as a teaching faculty member to my current colleagues, numerous people have responded with "so you'll be their version of Justin?" and I aspire to be even half as influential as he has been here at Caltech.

Throughout my time in the lab, I have seen many fellow graduate students come and go. Regardless of the precise make-up of the lab, the Phillips lab has always been marked by its strong collaborative spirit and I have rarely had to work alone. Specifically, when I first joined the lab, I was graciously guided by Stephanie Barnes, Nathan Belliveau, and Bill Ireland as I was ushered into the Sort-Seq part of the group. Though a daunting task, Bill and I worked together for many years to bring the Sort-Seq approach into the new era of Reg-Seq. With this new era came a new set of people working on the regulatory genome of *E. coli*: Tom Röscher and Scott Saunders have recently joined the lab and Manuel Razo-Mejia has switched gears slightly from his in-depth dissections of the *lac* operon to more genome-wide approaches, through the use of RNA-Seq. I have deeply enjoyed my time with them all as the "socialists," and while I am the first of our group to leave Caltech, I can't wait to see where our project ends up, even if it is from afar.

Other members of the lab, though I have not had the opportunity to work with them directly, have also been a source of immeasurable support. Griffin Chure has been an friend, ally, and font of knowledge throughout my time in lab. He is an inspiration not just as a scientist, but also as a human being. While we hardly spoke back when I was first rotating in the lab, I'm glad I've gotten to know Soichi Hirokawa better through our adventures to Woods Hole, New Zealand, and Korea. He helped me face some truly harrowing situations including the Tokyo subway system and a night with far too much soju. Lastly, the following members of the lab have been a source of support over the years, whether as administrators, office mates, or company on coffee breaks: Pamela Albertson, Rachel Banks, Celene Barrera, Kimberly Berry, Adam Catching, Ana Duarte, Tal Einav, Avi Flamholz, Helen Bermudez Foley, Vahe Galstyan, Jonathan Gross, Zofii Kaczmarek, Heun Jin Lee, Gita Mahmoudabadi, Niko McCarty, Muir Morrison, Rebecca Rousseau, Gabe Salmon, and Franz Weinert.

Outside of the lab, there are numerous people who have supported me. Primarily, the self proclaimed 'Broads of Broad': Grace Chow, Annisa Dea, Sarah Gillespie, Heidi Klumpe, Christina Su, Lynn Yi, and Shinae Yoon. They have all been a source of inspiration and served to make my time as a woman at Caltech less isolating.

Outside of Caltech, Misha Vysotskiy has been my best friend throughout grad school and well before. We've faced every academic stage together, starting with taking CS70 together at Harvey Mudd College in 2012. From there, we were Mathematical and Computational Biology majors, faced the graduation school application process, and even traveled to some interviews together. And we're still standing after all this time.

On a personal side, I want to thank those who have been with me long before my time at Caltech began: my parents. Whether helping me learn to read with flashcards, testing me on my spelling words on the way to school, or looking over my math homework, they have always been committed to me and my academic success. Their endless sacrifice made sure I endured even through a childhood riddled with illness and other hiccups. I'm incredibly privileged and lucky to have 'made it', and I know it would not have been possible without their commitment to me.

Lastly, I am so incredibly grateful to have spent the last 8+ years with my partner Tobin Ivy, and look forward to many, many more. During our time at Caltech, he has evolved from my boyfriend to my roommate to my fiancé. I can't wait to start the next chapter of our lives, building a life and a home together in Colorado.

ABSTRACT

Advances in DNA sequencing have revolutionized our ability to read genomes. However, even in the most well-studied of organisms, the bacterium *Escherichia coli*, for $\approx 65\%$ of promoters we remain ignorant of their regulation. Until we crack this regulatory Rosetta Stone, efforts to read and write genomes will remain haphazard. We introduce a new method, Reg-Seq, that links massively-parallel reporter assays with mass spectrometry to produce a base pair resolution dissection of more than 100 *E. coli* promoters in 12 growth conditions. We demonstrate that the method recapitulates known regulatory information. Then, we examine regulatory architectures for more than 80 promoters which previously had no known regulatory information. In many cases, we also identify which transcription factors mediate their regulation. This method clears a path for highly multiplexed investigations of the regulatory genome of model organisms, with the potential of moving to an array of microbes of ecological and medical relevance.

PUBLISHED CONTENT AND CONTRIBUTIONS

Ireland, W. T., S. M. Beeler, E. Flores-Bautista, N. S. McCarty, T. Röschinger, N. M. Belliveau, M. J. Sweredoski, A. Moradian, J. B. Kinney, and R. Phillips (2020). “Deciphering the regulatory genome of *Escherichia coli*, one hundred promoters at a time”. In: *eLife*. ISSN: 2050084X. DOI: 10.7554/eLife.55308. eprint: 2001.07396.

S.M.B helped design, optimize, and conduct experiments, and helped write and create figures for the manuscript.

TABLE OF CONTENTS

Acknowledgements	iii
Abstract	v
Published Content and Contributions	vi
Table of Contents	vi
List of Illustrations	viii
List of Tables	x
Chapter I: Introduction: On a quantitative understanding of gene regulation	1
1.1 Introduction	1
1.2 The discovery of molecular adaptation	1
1.3 The molecular players of gene regulation: the <i>lac</i> operon as an example	3
1.4 Statistical mechanics of gene regulation	5
1.5 On our regulatory ignorance	10
1.6 From dissection to exploration	12
1.7 A primer on mutual information	15
1.8 In conclusion: where we go from here	18
Chapter II: Deciphering the regulatory genome of <i>Escherichia coli</i> , one hundred promoters at a time	21
2.1 Abstract	21
2.2 Introduction	21
2.3 Results	25
2.4 Discussion	51
2.5 Methods	53
2.6 Supplementary information: Extended details of experimental design	61
2.7 Supplementary information: Validating Reg-Seq against previous methods and results	64
2.8 Supplementary information: Extended details of analysis methods	75
2.9 Supplementary information: Additional Results	95
2.10 Supplementary information: Key Resource Table	104
Chapter III: Quantitative dissection of a single promoter using RNA-Seq	111
3.1 Motivation	111
3.2 Preliminary results	112
3.3 Supplementary information: library content and design	113
3.4 Supplementary information: ORBIT cloning protocol	115
Chapter IV: Concluding Thoughts and Future Directions	118
4.1 Progress	118
4.2 Future goals	119
4.3 Outstanding challenges	120
Bibliography	124

LIST OF ILLUSTRATIONS

<i>Number</i>	<i>Page</i>
1.1 Examples of diauxic growth.	2
1.2 Regulation of the <i>lac</i> operon.	4
1.3 Microstates of RNA polymerase binding to DNA.	6
1.4 Pictorial representation of p_{bound}	7
1.5 States and weights for polymerase binding.	8
1.6 States and weights for simple repression.	9
1.7 Theory meets experiment for simple repression.	10
1.8 Regulatory ignorance across the domains of life.	11
1.9 Evidence of gene regulation in un-annotated genes.	12
1.10 The Sort-Seq protocol.	13
1.11 Expression shift for the mutagenized <i>lac</i> promoter.	14
1.12 An example of mutual information on a piece of text.	16
1.13 An example of mutual information between DNA sequence and gene expression.	18
1.14 All regulatory architectures uncovered in this thesis.	19
2.1 The <i>E. coli</i> regulatory genome.	26
2.2 Schematic of the Reg-Seq procedure as used to recover a repressor binding site.	27
2.3 A summary of four direct comparisons of measurements from Sort- Seq and Reg-Seq.	29
2.4 All regulatory architectures uncovered in this study.	34
2.5 Examples of the insight gained by Reg-Seq in the context of promoters with no previously known regulatory information.	36
2.6 A summary of regulatory architectures discovered in this study.	38
2.7 GlpR as a widely-acting regulator.	47
2.8 FNR as a global regulator.	48
2.9 Inspection of a genetic circuit.	49
2.10 Representative view of the interactive figure that is available online.	50
2.11 Procedure to identify binding site regions automatically.	60
2.12 Schematic of the genetic construct used in this study.	62
2.13 Mock data comparing Sort-Seq and Reg-Seq sequence logo values.	67

2.14	A visual comparison of the literature binding sites and the extent of the binding sites discovered by our algorithmic approach.	75
2.15	A visual display of the results of the TOMTOM motif comparison between the discovered binding sites and known sequence motifs from RegulonDB and our prior Sort-Seq experiment	76
2.16	Pearson correlation as a function of the number of unique DNA sequences.	85
2.17	Motif comparison using TOMTOM for the two PhoP binding sites in the <i>ybjX</i> promoter.	94
2.18	Two cases in which we see transcription factor binding sites that we have found to regulate both of the two divergently transcribed genes. .	95
2.19	A comparison of the types of architectures found in RegulonDB to the architectures with newly discovered binding sites found in the Reg-Seq study.	97
3.1	Theory meets experiment for simple repression.	112
3.2	Barcode coverage for the three wildtype operator sequences.	113
3.3	Operator coverage of the O1 library.	114

LIST OF TABLES

<i>Number</i>	<i>Page</i>
2.1 All promoters examined in this study, categorized according to type of regulatory architecture.	39
2.2 All genes investigated in this study categorized according to their regulatory architecture	45
2.3 All growth conditions used in the Reg-Seq study.	65
2.4 A suite of experimentally validated and high-evidence binding sites used to test our automated binding site finding algorithm.	69
2.5 The results of the comparison between experimentally verified, high evidence binding sites and Reg-Seq binding sites.	71
2.6 Example dataset of 4 nucleotide sequences, and the corresponding counts from the plasmid library and mRNAs.	77
2.7 Global, absolute quantification for most transcription factors identified in this study.	86
2.8 Example energy matrix.	89
2.9 Example dataset with energy predictions.	90
2.10 Scaling factors to convert arbitrary units to absolute units in $k_B T$. . .	92
2.11 Key Resource Table.	110

Chapter 1

INTRODUCTION: ON A QUANTITATIVE UNDERSTANDING OF GENE REGULATION

1.1 Introduction

Adaptation is nearly synonymous with being alive. The commonly used adage ‘life finds a way’ hints at the universality of adaptation within the living world. The concept of adaptation should be familiar from our day-to-day lives. As an example, our eyes are able to adjust from broad daylight to a dark room in a matter of minutes through the dilation of our pupils, permitting more light to enter our eyes. In this way, we adapt to our environment in a way that makes it more suitable for our survival. It should come as no surprise that adaptation such as this occurs across all domains of life, although the exact mechanism of adaptation may be qualitatively different than the example given here. In the broadest of strokes, this thesis is about adaptation, specifically how the bacterium *Escherichia coli* enacts adaptation at the molecular level. Plainly, this could be couched in the language of whether a given gene is either “on” or “off” in a given environmental condition. However, as I will argue here, we can move well beyond this qualitative language, to a more quantitatively rigorous and precise formulation, one that will give us a deeper understanding of how cells adapt.

1.2 The discovery of molecular adaptation

When exploring the fascinating ways in which bacteria and namely *Escherichia coli* adapt to their surrounding environments, we must acknowledge that we are standing on the shoulders of giants and give nod to the foundational work that began nearly 80 years ago by Jacques Monod. From a now seemingly simple experiment of providing bacteria with two sugars (glucose and arabinose for example), Monod discovered an interesting pattern of growth, resulting in the famed diauxic growth curves (Figure 1.1). Such growth curves are distinguished by two distinct growth phases, one where the preferred sugar (glucose in this case) is metabolized, followed by growth on the secondary sugar. The transition between the two growth phases can clearly be seen as a distinct secession of growth, a period of “adaptation”. The mystery presented by these growth curves is precisely what is occurring in the cells during this period of adaptation.

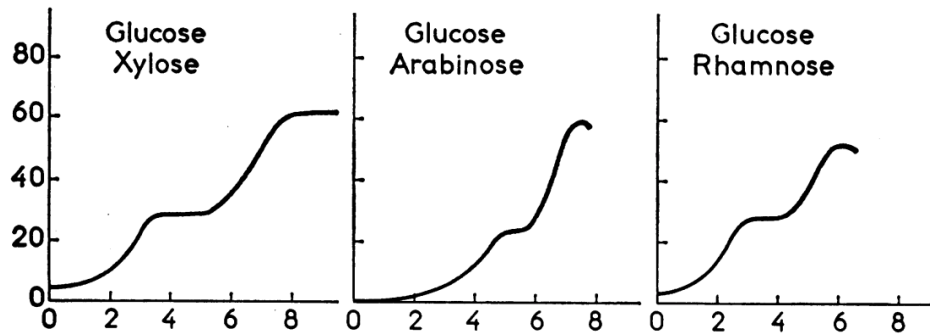


Figure 1.1: Examples of diauxic growth from Monod's thesis (1941), as reproduced in Monod, 1966. The x-axis is time in hours and the y-axis is optical density (arbitrary units), a metric of cell growth.

As Monod recalls in reflecting on his initial discovery, his colleague André Lwoff at the time suggested that it might have something to do with “enzymatic adaptation” (Monod, 1966), the idea that some protein already present in the cell responds to the changes in relative sugar concentrations. And thus the notion of adaptation on the molecular level was created. While the idea that enzymes themselves can change in response to surroundings (e.g. activity changed in response to the binding of the ligand), the primary cause of the lag in growth was due not to protein response alone, but to the need for a new suite of genes to be expressed and produced in response to the change in sugar. Monod's original discovery led to a decades-long journey of teasing apart how such gene regulation is enacted.

Next we will discuss the broad strokes nature of gene regulation, the way in which the expression of given genes can be modulated by the binding of proteins to DNA known as transcription factors. By way of example, we will specifically consider the extensively studied *lac* operon here, but many other genes have undergone similar dissection.

A note on adaptation

As a quick aside, I want to riff on an alternate meaning of adaptation. While thus far we have discussed what could be referred to as *physiological adaptation*, the perhaps more common use of the word adaptation is within the context of *evolutionary adaptation*. Examples of adaptations in this context would be the formation of webbed feet to aid in swimming or the use of prehensile tails for effective climbing in trees. The timescales required for these incredible feats to have evolved are nearly irrelevant for our present discussion regarding *molecular*

adaptation and the ways in which cells respond in real time to their surroundings. While the concept of evolutionary adaptation is not the primary focus of this thesis, it is important to note that the molecular mechanisms that are in place to permit cells to readily adapt to their surroundings are themselves subject to natural selection. As such, the ability to adapt is itself an adaptation, but for now, we focus our efforts on how cells enact their molecular adaptation rather than how such adaptations arose in the first place.

1.3 The molecular players of gene regulation: the *lac* operon as an example

With decades of painstaking experiments, the field of molecular genetics was able to make sense of the diauxic curves that Monod first discovered in the 1940s. In the specific case of cells provided with glucose and lactose, the molecular mechanisms of diauxic growth are illustrated in Figure 1.2. While specifically these lactose metabolizing genes are the primary focus of this section and arguably the most well-studied gene in *E. coli*, the mechanisms discussed here have far broader reach than just this single set of genes in this single organism. Most notably, our primary focus will be on a suite of proteins known as transcription factors, which bind to DNA and accordingly influencing the level of transcription (i.e. gene expression). Within the class of transcription factors, these proteins either act as *activators* to increase transcription or *repressors* to prevent or lower transcription. As an aside, it is possible for a given transcription factor to act as both an activator and a repressor, a duality known as a ‘Janus molecule’. However, this switch from activation to repression is mediated by some environmental cue, still making it reasonable to break transcription factors into these two discrete groups at least when considering a given gene and a given environment.

By way of example, we will work through the logic of the *lac* operon, illustrating the roles that activators and repressors can play in mediating the level of gene expression. Conveniently, the *lac* operon has one activator and one repressor that modulate its expression, making it a useful example to work through. As a way to understand the regulatory logic of this operon, it is important to remember that the *lac* operon encodes a number of genes involved in lactose metabolism, thus it make sense that the cell would only ‘want’ these genes to be turned on in the presence of lactose. Indeed, we can see that this is precisely how the regulation of the *lac* operon is enacted, with the repressor bound when there is no lactose around (Figure 1.2). With this set up, these lactose metabolizing genes will not be needlessly produced when their target substrate is not present. Conversely, as

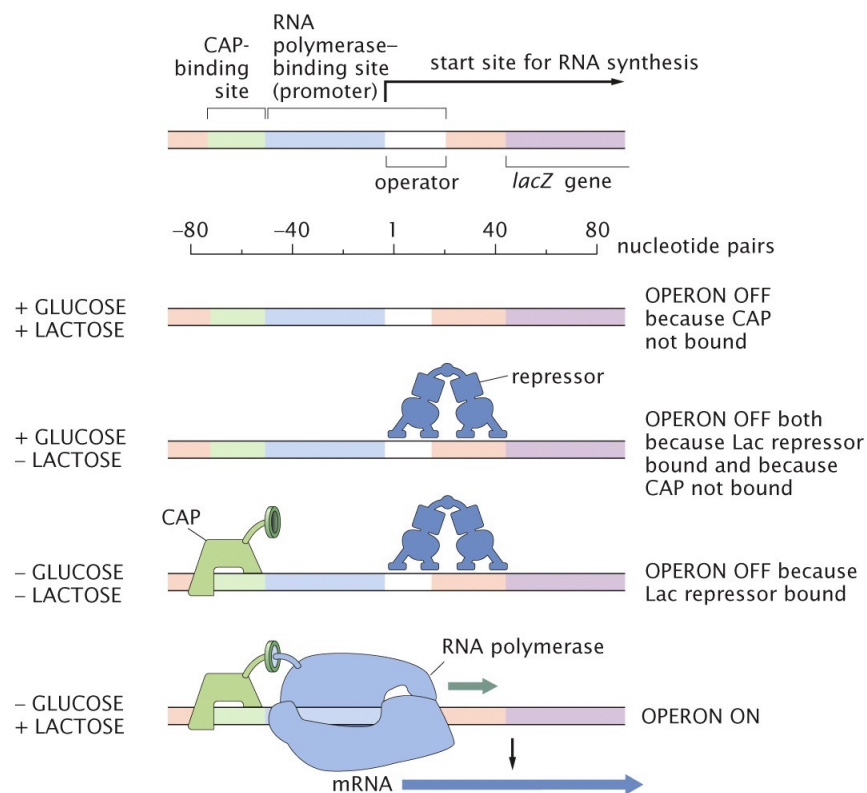


Figure 1.2: Regulation of the *lac* operon. The top schematic illustrates the key DNA regions of the *lac* operon. The GLUCOSE and LACTOSE labels on the left indicate the four possible environmental conditions in the presences (+) or absence (-) of these sugars, with the resulting regulatory state on the right. From this, we can see that the *lac* genes are only actively transcribed in the presence of lactose and the absence of glucose. Figure reproduced from *Physical Biology of the Cell*.

hinted by the shape for the diauxic growth curves, there are some sugars that are preferred over others. Specifically, glucose is the most preferred sugar as it can be immediately consumed through the citric acid cycle, while other sugars are only metabolized as an alternative. With this stipulation in mind, the cell would also ‘want’ to not express the lactose metabolizing genes when there is a perfectly more suitable sugar around such as glucose. Examining Figure 1.2, we can see that such logic is encoded molecularly through the use of the catabolite activator protein (known as CAP). That is, only in the absence of glucose does CAP serve to promote transcription by binding upstream of the *lac* promoter. Together through the action of the LacI repressor and the CAP activator, the regulation of these genes are effectively controlled as an AND gate, where *both* the presence of lactose *and* the absence of glucose are required for transcription.

While I am glossing over decades worth of hard-earned results as summarized by a single figure, it suffices to say that it *is* in fact possible to gain such a detailed understanding of how a given gene is regulated, i.e. which transcription factors are binding, where they bind, and whether they act as an activator or repressor. The next section will delve into what we can do with such a regulatory model in hand.

1.4 Statistical mechanics of gene regulation

As will be a common theme throughout this thesis, we will argue for moving beyond a qualitative understanding of gene regulation, as typified by the ‘cartoon’ models, like those in Figure 1.2. Instead, we would like to be able to have a mathematical model in addition to our pictorial one. The reason is simple: data in molecular biology are becoming ever more quantitatively precise, and accordingly our hypotheses should be similarly precise. For this, we will rely on a physical framework known as *statistical mechanics*. This section provides a brief overview of how the tools of statistical mechanics can be brought to bear on gene regulation.

A key tenet of statistical mechanics is described by Boltzmann distribution which states that the probability of a given state occurring is

$$P_{\text{state}} = \frac{e^{-\frac{\epsilon_{\text{state}}}{k_B T}}}{\mathcal{Z}}, \quad (1.1)$$

where ϵ_{state} is the energy associated with the state, k_B is the Boltzmann constant, and T is the temperature, and \mathcal{Z} is known as the partition function, or the sum of the probabilities of all possible states. In words, this equates to lower energy states being more likely to occur and higher energy states becoming vanishingly less likely to occur, due to the exponent. We can make sense of this intuitively to explain why we don’t spontaneously begin levitating. The energy, specifically the potential energy mgh , associated with the “state” of levitating is simply too large for it to realistically ever occur for large masses m such as ourselves. Now we must contend with what exactly is meant by a ‘state’ in statistical mechanical sense, starting with an example of RNA polymerase (RNAP) binding to DNA.

A state is some condition that we are interested in assessing the frequency of, such as whether an RNAP is bound to a promoter of interest. An additional definition is that a microstate is simply one specific manifestation of a state of interest. As illustrated in the bottom panel of (Figure 1.3), there are many ways to realize the binding of RNAP to DNA, each one its own microstate. Specifically, if we discretize

the genome, which is not an unreasonable assumption given that RNAP binds in register with specific basepairs, there end up being a total of N_{NS} nonspecific binding sites for the P polymerases to find themselves. The task at hand is to enumerate all these possible microstates, which can be defined as

$$W(P, N_{NS}) = \binom{N_{NS}}{P} = \frac{N_{NS}!}{P!(N_{NS} - P)!}, \quad (1.2)$$

where we use the notation that W stands for the number of microstates.

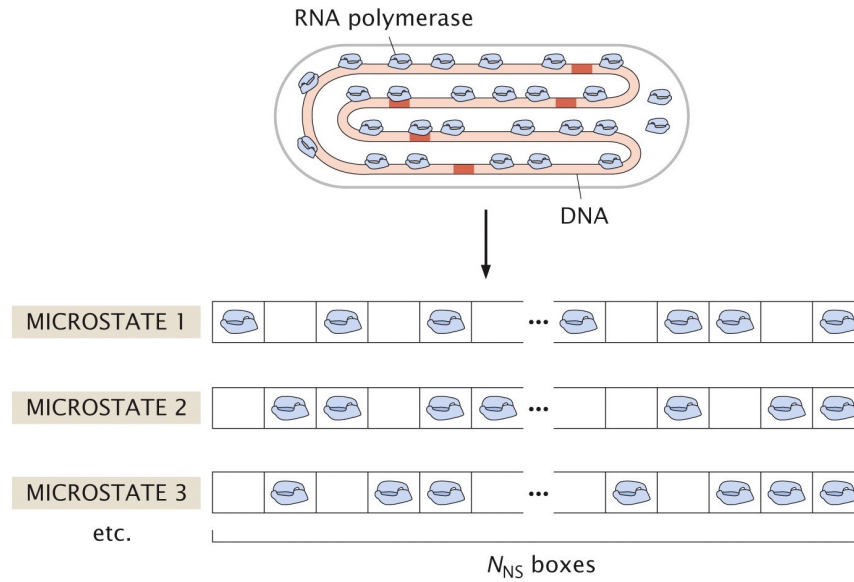


Figure 1.3: Microstates of RNA polymerase binding to DNA. The top panel schematizes the cell's pool of RNA polymerase (RNAP) as bound in various location along the length of the bacterial (circular) genome. The bottom panel illustrates three specific realizations (i.e. microstates) of the ways in which the RNAP may allocate themselves within the N_{NS} 'boxes' or basepairs of the genome. Figure reproduced from *Physical Biology of the Cell*.

With the enumeration of the states, along with Boltzmann distribution (Equation 1.1), we are poised to assess the probability of a promoter being bound by RNAP. That is, the probability of a state occurring is a function of both its associated energy (as prescribed by the Boltzmann distribution) *and* the number of ways in which a given state can occur, known as the *multiplicity*. Put together, we end up with the partial partition function for nonspecific binding as

$$\mathcal{Z}_{NS}(P, N_{NS}) = \underbrace{\frac{N_{NS}!}{P!(N_{NS} - P)!}}_{\text{multiplicity}} \times \underbrace{e^{-\beta P \epsilon_{pd}^{NS}}}_{\text{Boltzmann factor}}, \quad (1.3)$$

where we have introduced the simplifying notation that $\beta = 1/k_B T$ and have defined ϵ_{pd}^{NS} as the binding energy of polymerase to DNA at a nonspecific location. What Equation 1.3 describes is the probabilistic weight associated with all P polymerases being bound nonspecifically. However, if we are interesting in when a gene is being actively expressed, we would want to assess the probability that a polymerase is in fact bound to the promoter we are interested in, as schematized in Figure 1.4. That is, we want to compare the probabilistic weight of the promoter being bound relative to all possible states (promoter bound or unbound). For obtaining the partial partition function for polymerase being bound to the promoter, this amounts to effectively taking one polymerase out of circulation and placing the remaining $P - 1$ polymerases on the N_{NS} genome positions. This results in the following partial partition function for when a polymerase is bound to the promoter of interest:

$$\mathcal{Z}_{NS}(P - 1, N_{NS}) = \underbrace{\frac{N_{NS}!}{(P - 1)!(N_{NS} - (P - 1))!}}_{\text{multiplicity}} \times \underbrace{e^{-\beta(P-1)\epsilon_{pd}^{NS}} e^{-\beta\epsilon_{pd}^S}}_{\text{Boltzmann factor}}. \quad (1.4)$$

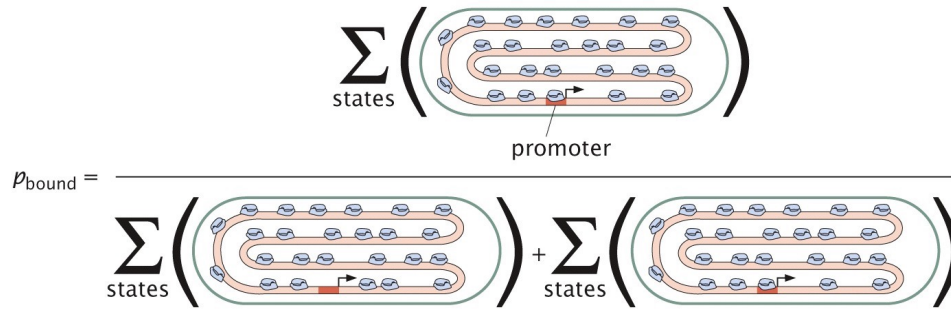


Figure 1.4: Pictorial representation of p_{bound} . The numerator is the sum of all the microstates in which the promoter of interest is bound by polymerase. The denominator is the sum all all states (i.e. those where the promoter is occupied and those where the promoter is unoccupied). Figure reproduced from *Physical Biology of the Cell*.

Note how the multiplicity is described by placing $P - 1$ polymerases, and the Boltzmann factor now also has $P - 1$ instances of nonspecific binding in addition to one

instance of specific binding, with energy ϵ_{pd}^S . These computations of the probabilistic weights are illustrated in Figure 1.5. We can simplify the multiplicities slightly by making the approximation $N_{NS}!/(N_{NS}-P)! \approx (N_{NS})^P$, with the reasonable assumption that $P \ll N_{NS}$. This approximation leaves us with the following weights:

$$\mathcal{Z}_{NS}(P, N_{NS}) = \frac{(N_{NS})^P}{P!} e^{-\beta P \epsilon_{pd}^{NS}}, \quad (1.5)$$

and

$$\mathcal{Z}_{NS}(P-1, N_{NS}) = \frac{(N_{NS})^{P-1}}{(P-1)!} e^{-\beta(P-1)\epsilon_{pd}^{NS}} e^{-\beta \epsilon_{pd}^S}. \quad (1.6)$$

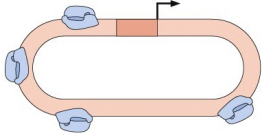
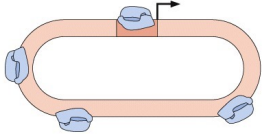
STATE	ENERGY	MULTIPLICITY	WEIGHT (MULTIPLICITY \times BOLTZMANN WEIGHT)
	$P \epsilon_{pd}^{NS}$	$\frac{N_{NS}!}{P! (N_{NS}-P)!} = \frac{(N_{NS})^P}{P!}$	$\frac{(N_{NS})^P}{P!} e^{-P \epsilon_{pd}^{NS} / k_B T}$
	$(P-1) \epsilon_{pd}^{NS} + \epsilon_{pd}^S$	$\frac{N_{NS}!}{(P-1)! [N_{NS}-(P-1)]!} = \frac{(N_{NS})^{P-1}}{(P-1)!}$	$\frac{(N_{NS})^{P-1}}{(P-1)!} e^{-(P-1) \epsilon_{pd}^{NS} / k_B T} e^{-\epsilon_{pd}^S / k_B T}$

Figure 1.5: States and weights for polymerase binding. The top panel works through the computation for the Boltzmann weight for the state of the promoter of interest being unoccupied by polymerase. By contrast, the bottom panel computes the weight for an occupied promoter. Figure reproduced from *Physical Biology of the Cell*.

At long last we are equipped to assess the probability that the promoter is in fact bound by polymerase, a state we will use as a proxy for gene expression. With Figure 1.4 as a visual aid for how to compute p_{bound} , we arrive at

$$p_{\text{bound}} = \frac{\frac{(N_{NS})^{P-1}}{(P-1)!} e^{-\beta(P-1)\epsilon_{pd}^{NS}} e^{-\beta \epsilon_{pd}^S}}{\frac{(N_{NS})^P}{P!} e^{-\beta P \epsilon_{pd}^{NS}} + \frac{(N_{NS})^{P-1}}{(P-1)!} e^{-\beta(P-1)\epsilon_{pd}^{NS}} e^{-\beta \epsilon_{pd}^S}}, \quad (1.7)$$

which is fairly daunting at first sight, but many values cancel out, leaving us with

$$p_{\text{bound}} = \frac{1}{1 + \frac{N_{NS}}{P} e^{\beta \Delta \epsilon_{pd}}}, \quad (1.8)$$

where $\Delta \epsilon_{pd}$ is defined as the difference between specific and nonspecific binding.

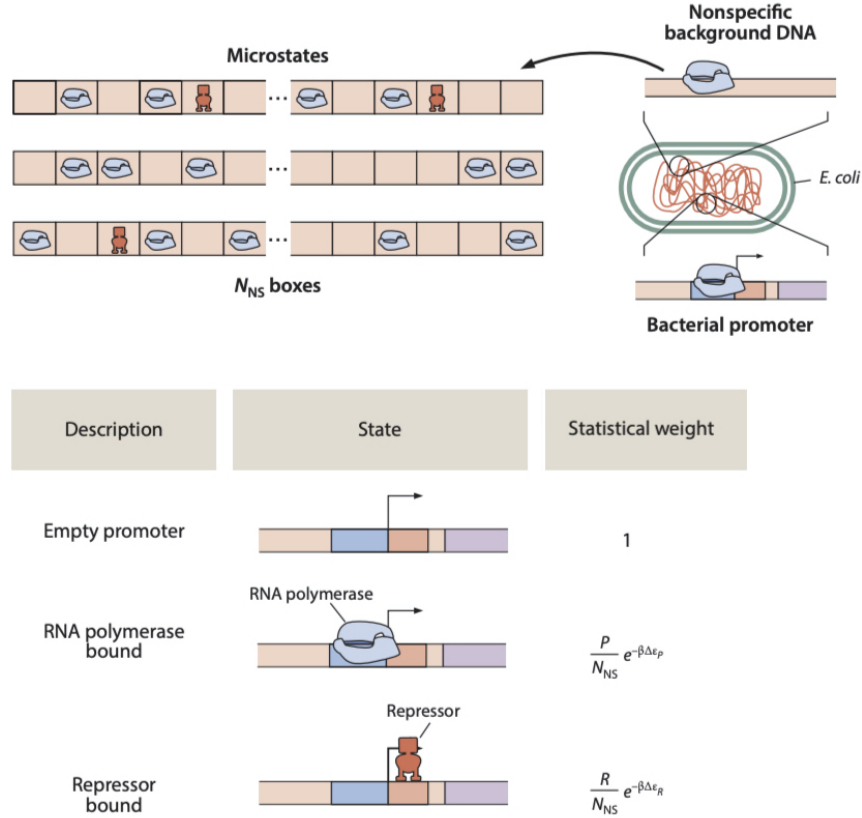


Figure 1.6: States and weights for simple repression. Figure adapted from Phillips et al., 2019.

While all this effort for modeling constitutive gene expression may seem rather arduous, there is great utility in this statistical mechanical protocol outlined here. If we wish to add the action of some transcription factor binding, say a repressor, it is actually quite simple to do so. The derivation we went through here applies by analogy to any protein binding to DNA. It is precisely the regulation enacted by a single repressor (a motif known as simple repression) that was the focus of work done by Brewster et al., 2014. While we won't go through the whole derivation again, we can use the same approaches outlined here for the constitutive promoter to arrive at the states and weights as shown in Figure 1.6.

It is with these states and weights in hand, we can now formulate a concrete mathematical prediction of how gene expression should change as a result of increasing repressor counts. It is precisely this theory-experiment dialogue that was conducted to much avail by Brewster et al., 2014, as shown in Figure 1.7. Such experiments serve to give us the sense that these statistical mechanical models of gene expression do actually fare well in describing the data. These careful mathematical models require knowing the precise regulatory structure of the promoter of interest, which works well for the thoroughly studied *lac* promoter. As we will see in the following section, however, there remain many genes for which such a treatment is not yet possible.

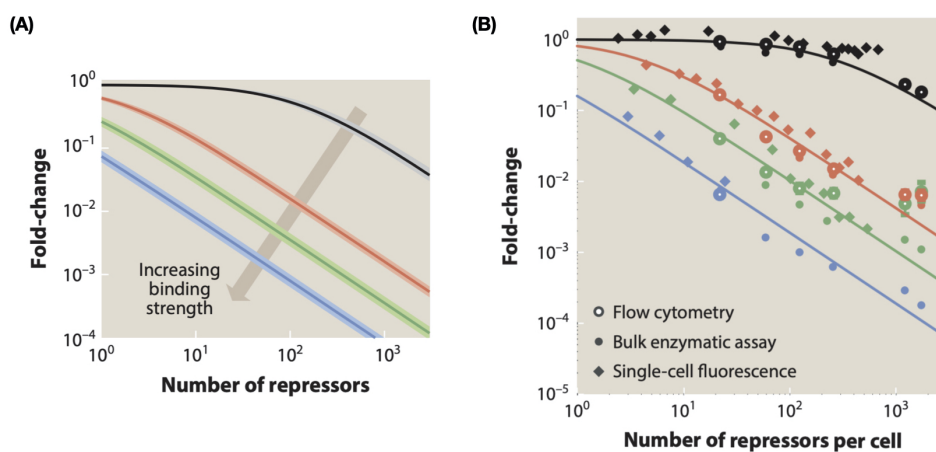


Figure 1.7: Theory meets experiment for simple repression. (A) Shows the prediction of how gene expression should change with increasing number of repressors, for four different binding sites. (B) Shows how the various data land relative to these predictions. Figure adapted from Phillips et al., 2019.

1.5 On our regulatory ignorance

Now that we have a cursory sense of the ways in which genes have been shown to be regulated and how we might mathematically model them, we come to one of the primary motivations of the work of this thesis: despite how much effort has been put into understanding how genes are regulated, especially in *E. coli*, there still remain many genes for which we know nothing regarding how they are regulated. Figure 1.8 (A) concisely illustrates the extent to which we remain ignorant of regulation even in the best case scenario of *E. coli*. Prior to the work conducted in this thesis, nearly two-thirds of all operons had no annotated gene regulation (i.e. any transcription factor binding sites), as annotated on RegulonDB Santos-Zavaleta

et al., 2019. (It's important to note that this value stated here includes the work done by Ireland et al., 2020, and that the number of genes with no known regulation was actually even greater prior to the work done in this thesis, as can be seen in Figure 2.1.) Unfortunately, as the remaining panel of Figure 1.8 reveal, the status of our regulatory ignorance only becomes worse as we move to higher organisms.

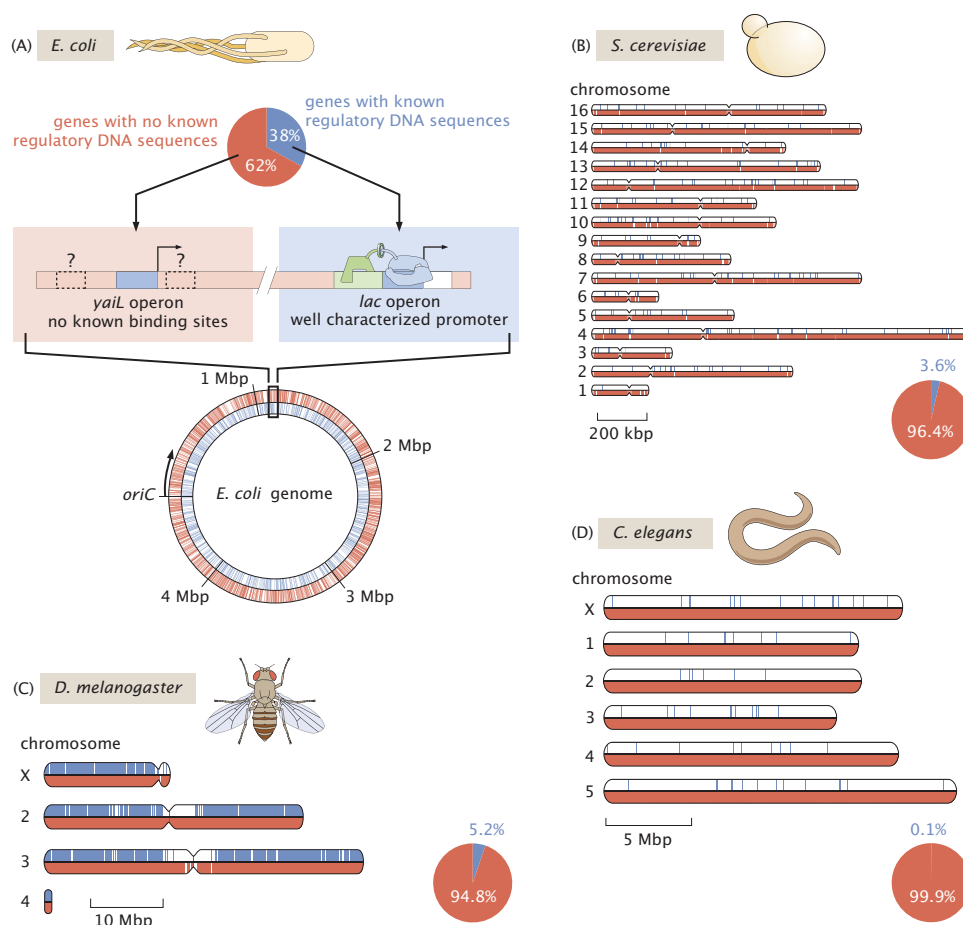


Figure 1.8: Regulatory ignorance across the domains of life. This schematized view of several genomes, showing each gene for which there is any known regulation (blue dashes) as opposed to those for which there was no known regulation (red dashes). This schematic shows our level of regulatory ignorance across (A) *E. coli*, (B) the budding yeast, *Saccharomyces cerevisiae*, (C) the fruit fly *Drosophila melanogaster*, and (D) the nematode *Caenorhabditis elegans*.

The primary battle cry of this thesis is that, as seen with the *lac* operon, we need to know some basic facts about the regulation of a given operon before we can begin to conduct the careful quantitative dissections discussed here. Thankfully, previous work can shed light on where regulation may be occurring even if we don't yet know

the details of that regulation. One such set of key experiments were conducted by Schmidt et al., 2016, where they assessed the full *E. coli* proteome over 22 unique growth conditions. By examining proteins whose expression changes dramatically across growth conditions, we can gain insight into genes whose expression likely seems to be regulated (and thus is only turned on in one or few conditions). As Figure 1.9 illustrates, there are in fact many proteins whose expression is highly variable across these 22 growth conditions, and with respect to our mission to explore the currently unannotated genes, we are heartened to see that both genes with known (in blue) *and* no known (in red) regulation demonstrate variable gene expression. It is precisely these genes in red with high coefficient of variation that serve as ideal candidates for uncovering hitherto unexplored regulation. Precisely how we do achieve that goal is introduced in the following section and is the primary thrust of the remainder of this thesis.

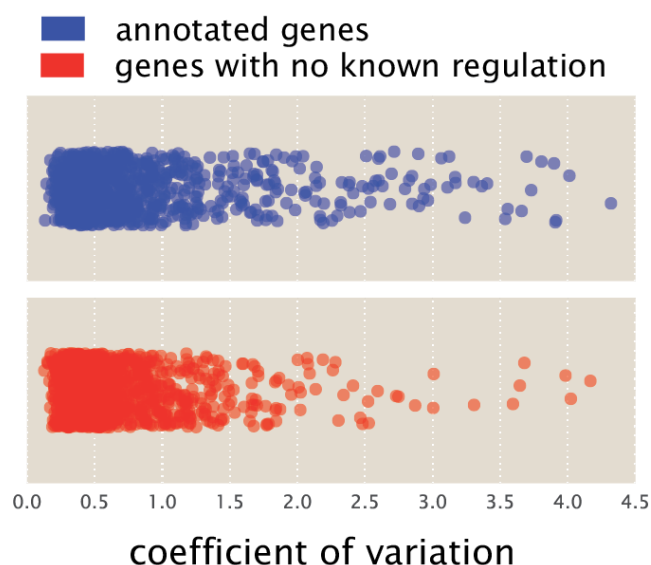


Figure 1.9: Evidence of gene regulation in unannotated genes. Each protein (categorized as having annotated regulation in blue or no known regulation in red) are plotted according to their coefficient of variation across the 22 growth conditions tested in Schmidt et al., 2016. Plots adapted from Belliveau et al., 2018.

1.6 From dissection to exploration

With the widespread issue of regulatory ignorance laid out clearly before us, we must now contend with how we can go about uncovering such previously unexplored genes. Work pioneered by Kinney et al., 2010 served to establish the Sort-Seq method, which Belliveau et al., 2018 later used to great avail to unveil

the regulation of some previously unexplored genes. The protocol is outline as in Figure 1.10. In brief, the scheme is to take a promoter region of interest and make a mutagenized library of promoter variants all driving expression of some fluorescent protein reporter gene. Using fluorescence-activated cell sorting (FACS), the initially heterogeneous population of cells can be separated into four distinct bins (Figure 1.10 (A)). The cells in each bin are then sequenced, allowing us to build up a picture of which mutations confer changes in fluorescence.

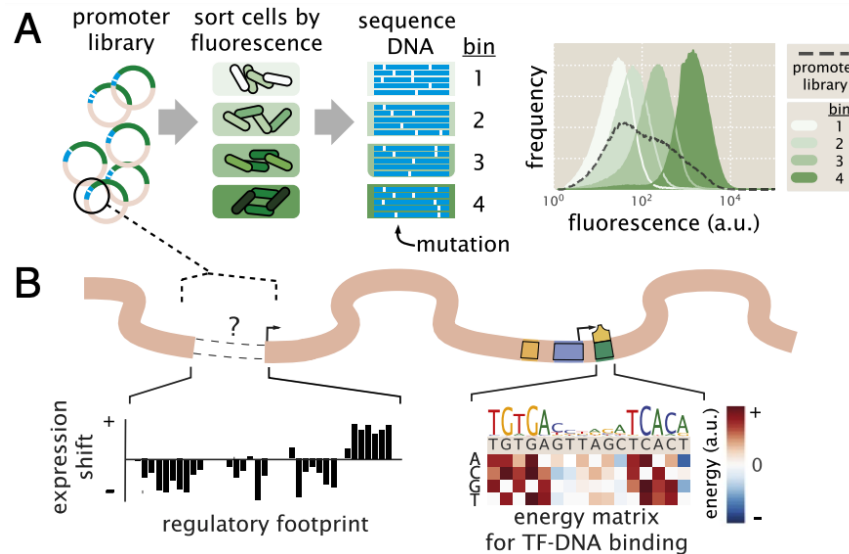


Figure 1.10: The Sort-Seq protocol. (A) By mutagenizing a promoter region of interest driving expression of some reporter protein, we can obtain cells with varying levels of fluorescence. These cells are sorted via flow cytometry into four bins and each bin is independently sequenced. The plot on the right shows the distribution of fluorescence of the four bins after having been sorted, demonstrating that the difference in gene expression is maintained within the disparate bin population. (B) With the sequencing information in hand, we can begin to assess the relative importance that each basepair has with respect to the level of gene expression. This formation is computed and displayed as expression shift plots (on the left) and energy matrices (on the right).

Specifically, along the length of the promoter region of interest, it is possible to assess whether a given basepair increases or decreases expression upon being mutated. Such information leads to an expression shift profile, as shown in left of Figure 1.10 (B). Such a plot gives a quick visual aid as to where transcription factors may be binding, as these regions are the most likely to have an impact on the level of expression. For example, if a given mutation disrupts the ability of a repressor to bind, we would expect such a variant to have higher gene expression than normal.

However we can also take a more detailed look at a given binding, as depicted by an energy matrix (right panel, Figure 1.10 (B)). Using the tools of statistical mechanics as outlined in Section 1.4, it is possible to directly connect the changes in gene expression to a binding energy. In this way, we are able to determine not just which basepairs are involved in regulating expression, but we can also concretely predict what effect various mutations will have.

With this technique in hand, it is essential to evaluate its ability to recover known regulation if it is to be of any use in uncovering our regulatory ignorance. By way of example, we once again return to the *lac* operon, whose regulation is well understood. Hearteningly, Belliveau et al., 2018 were in fact able to recover the known regulation for the promoter region when giving it the full Sort-Seq treatment, as revealed by the expression shift plot (Figure 1.11). Walking through these results, we can see that mutating the region where the *lacI* repressor binds causes the expression to go up on average. This makes sense as disrupting the binding of *lacI* will lead to a failure to repress the gene, ultimately causing gene expression to be higher. Conversely, we see that the regions where CAP and RNAP polymerase show the opposite effect, where mutation led to lower expression.

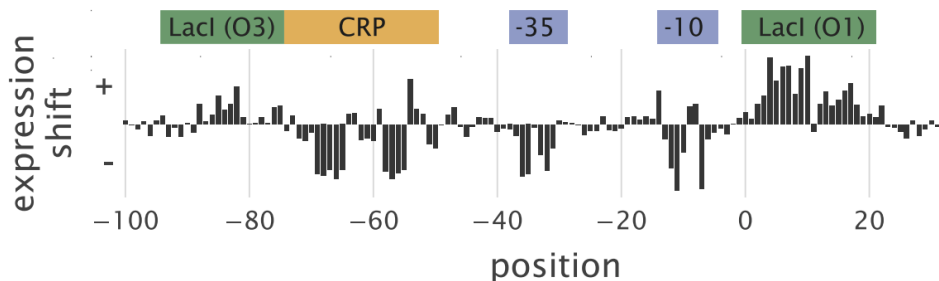


Figure 1.11: Expression shift for the mutagenized *lac* promoter. The plot shows the average effect of mutating a given position with respect to the resulting gene expression. The colored bars above the plot denote where the known binding sites are located. Data and figure from Belliveau et al., 2018.

These results encourage us that we can in fact use the Sort-Seq method to unveil gene regulation. In fact, the remainder of the work done by Belliveau et al., 2018 served to dissect two other promoters with known regulation and importantly four promoters with previously no known regulation. This incredibly important study served as an essential proof of concept as we embarked on the work discussed in this thesis. From here, we sought to expand the utility of Sort-Seq from dissecting

a single promoter at a time to being able to explore ten and even one hundred promoters at a time.

1.7 A primer on mutual information

Whether we are measuring fluorescence of a reporter protein or mRNA counts, we must now contend with how to make sense of the data in hand, specifically how to relate sequence identity to gene expression. For this we turn to the concept of *mutual information*, a metric by which we can understand how much information one variable provides about another. In this case, we would be interested in how much information a given DNA sequence gives us with respect to the resulting level of gene expression. Concretely, the mutual information between two discrete random variables X and Y is defined as

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right), \quad (1.9)$$

where $p(x, y)$ is the joint probability of x and y occurring. To gain more intuition into what this actually means, let's walk through an example that will be more familiar before delving into the case of gene expression. Let's take an iconic piece of text from the final sentence of Darwin's *On the Origin of Species*:

"There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved."

As an exercise, we can ask what the mutual information is between subsequent letters in this piece of text. That is, how much information does one letter give us about what letter is likely to follow? As prescribed in Equation 1.9, to compute the mutual information, we need both the joint probability of the two variables, $p(x, y)$ as well as their individual probabilities, $p(x)$ and $p(y)$, respectively. Figure 1.12 (A) illustrates these frequencies of each letter as found in this text. Intuitively, we can see that the letter 'e' is in fact the most common letter, and from here, we can begin to assess the joint probabilities as shown Figure 1.12 (B). Once again, we can make sense of the results shown here by noting the 't' followed by 'h' as well 'h'

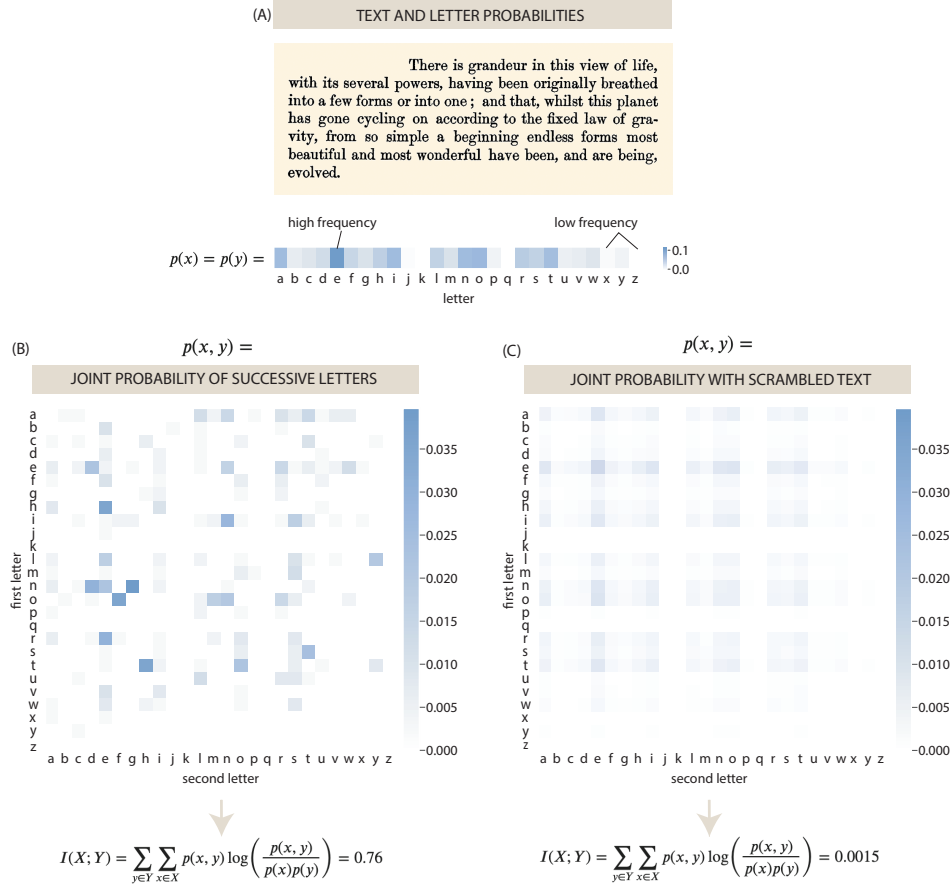


Figure 1.12: An example of mutual information on a piece of text. (A) The text from the final sentence of Darwin’s *On the Origin of Species*, along with the frequencies with which each letter appears. (B) The joint probabilities of two consecutive letters from the original text. (C) The joint probabilities of two consecutive letters when the text has been scrambled.

followed by ‘e’ are the most common, as expected by the common use of the word ‘the’.

Intuitively, we can see that there is in fact some information contained in this piece of text with regards to the identity of two consecutive letters. That is, if you were given a letter, you would be able to make a reasonable guess about which letter will follow (performing at least better than guessing randomly). To be more quantitatively precise, we can now plug in both the basal letter probabilities, both $p(x)$ and $p(y)$ in this case as shown Figure 1.12 (A) and the joint probabilities from Figure 1.12 (B) into Equation 1.9 to arrive at the total mutual information. Mechanistically, this entails looping through all possible letter combinations and

assessing both their joint probability $p(x, y)$ and their ‘expected’ probabilities of simply multiplying their independent frequencies together, $p(x) \times p(y)$. From this calculation, we arrive at value of 0.76 (Figure 1.12 (B)). However, to make sense of what this number means, we can by way of contrast scramble the text and repeat the procedure, as shown in Figure 1.12 (C). Such scrambled text might look something like this:

"pgfsfw ledeu e dtrxeiieui vh vtatbbetawo oo fear.di,n nor fsrmo ev ot
iaernvbws o,f lfmcats ebeeleruiim,tse rwohardntot dbo fctby r ndatla
f ndn in hhie ncei roocs nv n aaa, leml oes aselee uc.dhpdnltwan yeaai
htcghus xowfhgo w, iltenrhoerwaccsrau ftg;,h nedsr mmoo pon oeoer-
wioerid oit enohantosnrortgge hhhg enfttliem ylln, hwer le w,oa nmfi-
belhea ,a hfdaoo,lsna uiile fgmsatt pifaos,loavieeocareiefnefynenn etj
s hair pgftin ilteeyghiitrel hdh imsttblvsanrt,i sg o vdaiedtmman ise"

We can intuit that this piece of text has now sadly been rendered meaningless. We can see this more precisely in Figure 1.12 (C) that there are no letter combinations that are favored and ‘e’ becomes the most likely letter to follow regardless of the previous letter, as it is simply the most common letter. Lastly, we can quantify this by again making use of Equation 1.9. We now see that the mutual information is 0.0015, nearly 0, which would imply no information, as expected when the text has been scrambled and all the original meaning as been lost.

With what I hope is a more intuitive example in mind, we can finally return to the primary scientific question at hand: how much information does knowing a given DNA sequence give us about the resulting gene expression? As illustrated with a toy example in Figure 1.13, we can use the same exact approach to assess the joint probabilities between the identity of a given DNA nucleotide (A, T, C, or G) and the resulting gene’s expression (as measured by binning according the fluorescence level). In Figure 1.13, the first position can be seen to contain a high information, as knowing the identity of the nucleotide permits you to make a suitable guess as to which level of expression the DNA will promote. However, position 2 would have low information, as the joint probabilities are much more uniform and no single bin is particularly favored over another. By repeating this calculation over every nucleotide position along the length of DNA region of interest, we can begin to build a picture of which bases are most important for determining the level of expression. Such bases that are found to contain more information are thus heavily implicated

position 1					position 2					position n				
	bin1	bin2	bin3	bin4		bin1	bin2	bin3	bin4					
A	0.20	0.01	0.04	0.03	A	0.06	0.07	0.07	0.04					
T	0.01	0.08	0.16	0.02	T	0.04	0.09	0.03	0.08					
C	0.02	0.14	0.02	0.03	C	0.06	0.06	0.09	0.06					
G	0.02	0.02	0.03	0.17	G	0.09	0.03	0.06	0.07					

mutual information = 0.445 mutual information = 0.049

Figure 1.13: An example of mutual information between DNA sequence and gene expression. Tables show the joint probabilities between nucleotide identity along the rows and gene expression along the columns. The level of gene expression is discretized into bins, with increasing fluorescence indicated with intensity in green. Intensity in purple represents the value of the joint probabilities. Position 1 demonstrates high mutual information between the identity of the nucleotide and the level of gene expression (i.e. bin). By way of contrast, position 2 demonstrates lower mutual information as the joint probabilities are much more uniform.

in serving some regulatory role, such as serving as a binding site for a transcription factor. In the chapter that follows, it is precisely this approach that will be brought to bear on deciphering the yet-to-be understood regulatory regions of the *E. coli* genome.

1.8 In conclusion: where we go from here

With this introduction I hope I have impressed upon you two key themes: 1) the universality of adaptation, and specifically gene regulation as a lens through which to understand how cells adapt to their surroundings and 2) the need for a quantitative understanding of how gene regulation is enacted. However, as a first pass we need to know which transcription factors are even involved as well as where and how strongly they bind to a given promoter. With these motivating points in mind and a few “tricks of the trade” (i.e. statistical mechanics and mutual information) in hand, we are prepared to tackle the problem of our regulatory ignorance in *E. coli*. What follows is the magnum opus of my thesis, where we brought these tools to bear in deciphering a substantial chunk of the *E. coli* genome, one hundred genes in one set of experiments.

The results of my thesis can be concisely summarized in Figure 1.14. While the precise details of how these cartoon models were elucidated is left to Chapter 2,

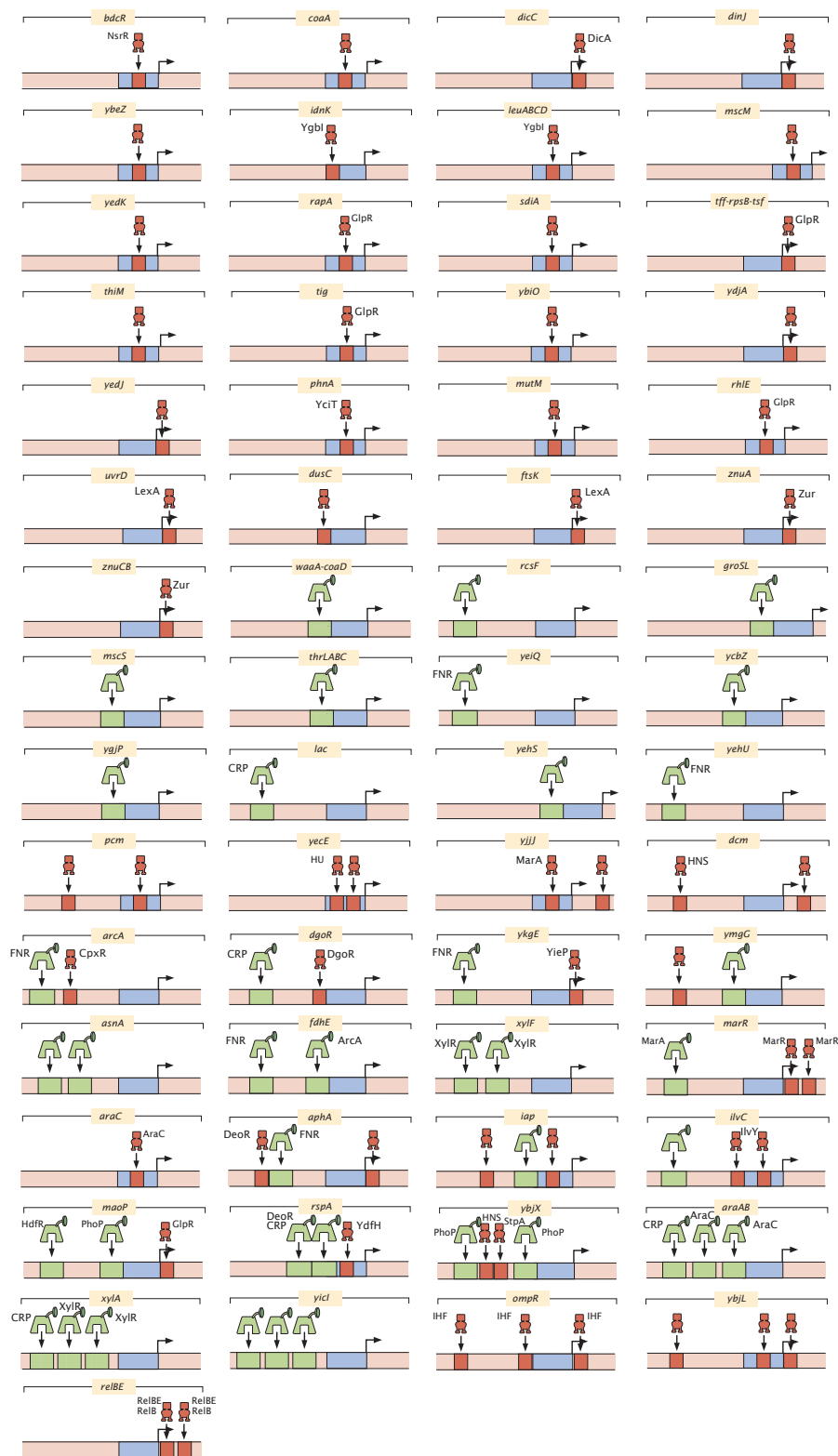


Figure 1.14: All regulatory architectures uncovered in this study. (Continued on the following page.)

Figure 1.14: For each regulated promoter, activators and their binding sites are labeled in green, repressors and their binding sites are labeled in red, and RNAP binding sites are labeled in blue. All cartoons are displayed with the transcription direction to the right. Only one RNAP site is depicted per promoter. Binding sites found for these promoters in the EcoCyc or RegulonDB databases are only depicted in these cartoons if the sites are within the 160 bp mutagenized region studied, and were detected by Reg-Seq.

this figure shows every binding site that was discovered across the 113 promoters explored here. It is important to note that the cartoon models shown here belie the precise quantitative backing that supports these results. That is, each binding site has its own information footprint and energy matrix as with the traditional Sort-Seq approach outlined in Figure 1.10. This means that not only do we know where transcription factors are binding, how they are regulating (i.e. as an activator or repressor), and in many cases the identity of the transcription factor, but we also know how strongly the given transcription factor binds. It is with this deep quantitative understanding of how these genes are regulated that we can make predictions as seen in Figure 1.7 and begin to test our understanding of how regulation is enacted well beyond what is illustrated by a cartoon alone.

The work discussed here has transformed the utility of Sort-Seq from being able to elucidate single genes to over a hundred genes at a time. With around 4000 genes in the *E. coli*, we can see a path to having the entire regulatory genome ‘solved’ within the coming years, hopefully radically transforming the view of Figure 1.8 (A). For now though, let’s dive into this first feat of tackling one hundred promoters.

*Chapter 2*DECIPHERING THE REGULATORY GENOME OF
ESCHERICHIA COLI, ONE HUNDRED PROMOTERS AT A
TIME

Published as:

Ireland, W. T., S. M. Beeler, E. Flores-Bautista, N. S. McCarty, T. Röschinger, N. M. Belliveau, M. J. Sweredoski, A. Moradian, J. B. Kinney, and R. Phillips (2020). “Deciphering the regulatory genome of *Escherichia coli*, one hundred promoters at a time”. In: *eLife*. ISSN: 2050084X. DOI: 10.7554/ELIFE.55308. arXiv: 2001.07396.

2.1 Abstract

Advances in DNA sequencing have revolutionized our ability to read genomes. However, even in the most well-studied of organisms, the bacterium *Escherichia coli*, for $\approx 65\%$ of promoters we remain ignorant of their regulation. Until we crack this regulatory Rosetta Stone, efforts to read and write genomes will remain haphazard. We introduce a new method, Reg-Seq, that links massively-parallel reporter assays with mass spectrometry to produce a base pair resolution dissection of more than 100 *E. coli* promoters in 12 growth conditions. We demonstrate that the method recapitulates known regulatory information. Then, we examine regulatory architectures for more than 80 promoters which previously had no known regulatory information. In many cases, we also identify which transcription factors mediate their regulation. This method clears a path for highly multiplexed investigations of the regulatory genome of model organisms, with the potential of moving to an array of microbes of ecological and medical relevance.

2.2 Introduction

DNA sequencing is as important to biology as the telescope is to astronomy. We are now living in the age of genomics, where DNA sequencing has become cheap and routine. However, despite these incredible advances, how all of this genomic information is regulated and deployed remains largely enigmatic. Organisms must respond to their environments through the regulation of genes. Genomic methods often provide a "parts list" but leave us uncertain about how those parts are used

creatively and constructively in space and time. Yet, we know that promoters apply all-important dynamic logical operations that control when and where genetic information is accessed. In this paper, we demonstrate how we can infer the logical and regulatory interactions that control bacterial decision making by tapping into the power of DNA sequencing as a biophysical tool. The method introduced here provides a framework for solving the problem of deciphering the regulatory genome by connecting perturbation and response, mapping information flow from individual nucleotides in a promoter sequence to downstream gene expression, determining how much information each promoter base pair carries about the level of gene expression.

The advent of RNA-Seq (Lister et al., 2008; Nagalakshmi et al., 2008; Mortazavi et al., 2008) launched a new era in which sequencing could be used as an experimental read-out of the biophysically interesting counts of mRNA, rather than simply as a tool for collecting ever more complete organismal genomes. The slew of ‘X’-Seq technologies that are available continues to expand at a dizzying pace, each serving their own creative and insightful role: RNA-Seq, ChIP-Seq, Tn-Seq, SELEX, 5C, etc. (Stuart and Satija, 2019). In contrast to whole genome screening sequencing approaches, such as Tn-Seq (Goodall et al., 2018) and ChIP-Seq (Gao et al., 2018), which give a coarse-grained view of gene essentiality and regulation respectively, another class of experiments known as massively-parallel reporter assays (MPRA) have been used to study gene expression in a variety of contexts (Patwardhan et al., 2009; Kinney et al., 2010; Sharon et al., 2012; Patwardhan et al., 2012; Melnikov et al., 2012; Kwasnieski et al., 2012; Fulco et al., 2019; Kinney and McCandlish, 2019). One elegant study relevant to the bacterial case of interest here by Kosuri et al., 2013 screened more than 10^4 combinations of promoter and ribosome binding sites (RBS) to assess their impact on gene expression levels. Even more recently, the same research group has utilized MPRA in sophisticated ways to search for regulated genes across the genome (Urtecho et al., 2019; Urtecho et al., 2020), in a way we see as being complementary to our own. While their approach yields a coarse-grained view of where regulation may be occurring, our approach yields a base-pair-by-base-pair view of how exactly that regulation is being enacted.

One of the most exciting X-Seq tools based on MPRA with broad biophysical reach is the Sort-Seq approach developed by Kinney et al., 2010. Sort-Seq uses fluorescence activated cell sorting (FACS) based on changes in the fluorescence due to mutated promoters combined with sequencing to identify the specific locations of

transcription factor binding in the genome. Importantly, it also provides a readout of how promoter sequences control the level of gene expression with single base-pair resolution. The results of such a massively-parallel reporter assay make it possible to build a biophysical model of gene regulation to uncover how previously uncharacterized promoters are regulated. In particular, high-resolution studies like those described here yield quantitative predictions about promoter organization and protein-DNA interactions (Kinney et al., 2010). This allows us to employ the tools of statistical physics to describe the input-output properties of each of these promoters which can be explored much further with in-depth experimental dissection like those done by Razo-Mejia et al., 2018 and Chure et al., 2019 and summarized in Phillips et al., 2019. In this sense, the Sort-Seq approach can provide a quantitative framework to not only discover and quantitatively dissect regulatory interactions at the promoter level, but also provides an interpretable scheme to design genetic circuits with a desired expression output (Barnes et al., 2019).

Earlier work from Belliveau et al., 2018 illustrated how Sort-Seq, used in conjunction with mass spectrometry, can be used to identify which transcription factors bind to a given binding site, thus enabling the mechanistic dissection of promoters which previously had no regulatory annotation. However, a crucial drawback of the approach of Belliveau et al., 2018 is that while it is high-throughput at the level of a single gene and the number of promoter variants it accesses, it was unable to readily tackle multiple genes at once. Even in one of biology's best understood organisms, the bacterium *Escherichia coli*, for more than 65% of its genes, we remain completely ignorant of how those genes are regulated (Santos-Zavaleta et al., 2019; Belliveau et al., 2018). If we hope to some day have a complete base pair resolution mapping of how genetic sequences relate to biological function, we must first be able to do so for the promoters of this "simple" organism.

What has been missing in uncovering the regulatory genome in organisms of all kinds is a large scale method for inferring genomic logic and regulation. Here, we replace the low-throughput, fluorescence-based Sort-Seq approach with a scalable, RNA-Seq based approach that makes it possible to attack many promoters at once. Accordingly, we refer to the entirety of our approach (MPRA, information footprints and energy matrices, and transcription factor identification) as Reg-Seq, which we employ here on over one hundred promoters. The concept of MPRA methods is to perturb promoter regions by mutating their sequences, and then to use next-generation sequencing (NGS) methods to read out how those mutations impact the

expression level of each promoter. (Patwardhan et al., 2009; Kinney et al., 2010; Sharon et al., 2012; Patwardhan et al., 2012; Melnikov et al., 2012; Kwasnieski et al., 2012; Fulco et al., 2019; Kinney and McCandlish, 2019). We generate a broad diversity of promoter sequences for each promoter of interest and use mutual information as a metric to measure the information flow from that distribution of sequences to gene expression. Thus, Reg-Seq is able to collect causal information about candidate regulatory sequences that is then complemented by techniques such as mass spectrometry, which allows us to find which transcription factors mediate the action of those newly discovered candidate regulatory sequences. Hence, Reg-Seq solves the causal problem of linking DNA sequence to regulatory logic and information flow.

To demonstrate our ability to perform Reg-Seq at scale, we report here our results for 113 *E. coli* genes, whose regulatory architectures (i.e. gene-by-gene distributions of transcription factor binding sites and identities of the transcription factors that bind those sites) were determined in parallel for multiple different growth conditions. Though we make substantial progress in mapping the regulatory information for a swath of *E. coli* genes in this study (the "regulome"), the field still remains limited in its understanding of which specific growth conditions, small molecules and metabolites (the allosterome) are responsible for altering the milieu of transcription factor activities (Lindsley and Rutter, 2006; Piazza et al., 2018; Huang et al., 2018). We hope to address this shortcoming in future studies by appealing to recent work on solving the "allosterome problem" (Piazza et al., 2018). By taking the Sort-Seq approach from a gene-by-gene method to a larger scale, more multiplexed approach, we can begin to piece together not just how individual promoters are regulated, but also the nature of gene-gene interactions by revealing how certain transcription factors serve to regulate multiple genes at once. This approach has the benefits of a high-throughput assay without sacrificing any of the resolution afforded by the previous gene-by-gene approach, allowing us to uncover the gene regulation of over 100 operons, with base-pair resolution, in one set of experiments.

The organization of the remainder of the paper is as follows. In the Results section, we benchmark Reg-Seq against our own earlier Sort-Seq experiments to show that the use of RNA-Seq as a readout of the expression of mutated promoters is equally reliable as the fluorescence-based approach. Additionally, we provide a global view of the discoveries that were made in our exploration of more than 100 promoters in *E. coli* using Reg-Seq. These results are described in summary form in the paper itself,

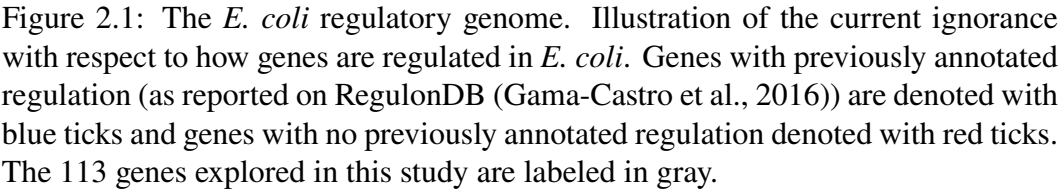
with a full online version of the results (www.rpgroup.caltech.edu/RegSeq/interactive) showing how different growth conditions elicit different regulatory responses. This section also follows the overarching view of our results by examining several biological stories that emerge from our data and serve as case studies in what has been revealed in our efforts to uncover the regulatory genome. The Discussion section summarizes the method and the current round of discoveries it has afforded with an eye to future applications to further elucidate the *E. coli* genome and open up the quantitative dissection of other non-model organisms. Lastly, in the Methods section and Appendices, we describe our methodology and the false positive and false negative rates of the method.

2.3 Results

Selection of genes and methodology

As shown in Figure 2.1, we have explored more than 100 genes from across the *E. coli* genome. Our choices were based on a number of factors (see Appendix 2.6 Section “Choosing target genes” for more details); namely, we wanted a subset of genes that served as a "gold standard" for which the hard work of generations of molecular biologists have yielded deep insights into their regulation. Our set of gold standard genes is *lacZYA*, *znuCB*, *znuA*, *ompR*, *araC*, *marR*, *relBE*, *dgoR*, *dicC*, *ftsK*, *xylA*, *xylF*, *rspA*, *dicA*, and *araAB*. By using Reg-Seq on these genes, we were able to demonstrate that this method recovers not only what was already known about binding sites of transcription factors for well-characterized promoters (Appendix 2.7, Figure 2.14), but also whether there are any important differences between the results of the methods presented here and the previous generation of experiments based on fluorescence and cell-sorting as a readout of gene expression (Kinney et al., 2010; Belliveau et al., 2018). These promoters of known regulatory architecture are complemented by an array of previously uncharacterized genes that we selected in part using data from a recent proteomic study, in which mass spectrometry was used to measure the copy number of different proteins in 22 distinct growth conditions (Schmidt et al., 2016). We selected genes that exhibited a wide variation in their copy number over the different growth conditions considered, reasoning that differential expression across growth conditions implies that those genes are under regulatory control.

As noted in the introduction, the original formulation of Reg-Seq, termed Sort-Seq, was based on the use of fluorescence activated cell sorting, one gene at a time, as a way to uncover putative binding sites for previously uncharacterized promoters



(Belliveau et al., 2018). As a result, as shown in Figure 2.2, we have formulated a second generation version that permits a high-throughput interrogation of the genome. A comparison between the Sort-Seq and Reg-Seq approaches on the same set of genes is shown in Figure 2.3. In the Reg-Seq approach, for each promoter interrogated, we generate a library of mutated variants and design each variant to express an mRNA with a unique sequence barcode. By counting the frequency of each expressed barcode using RNA-Seq, we can assess the differential expression

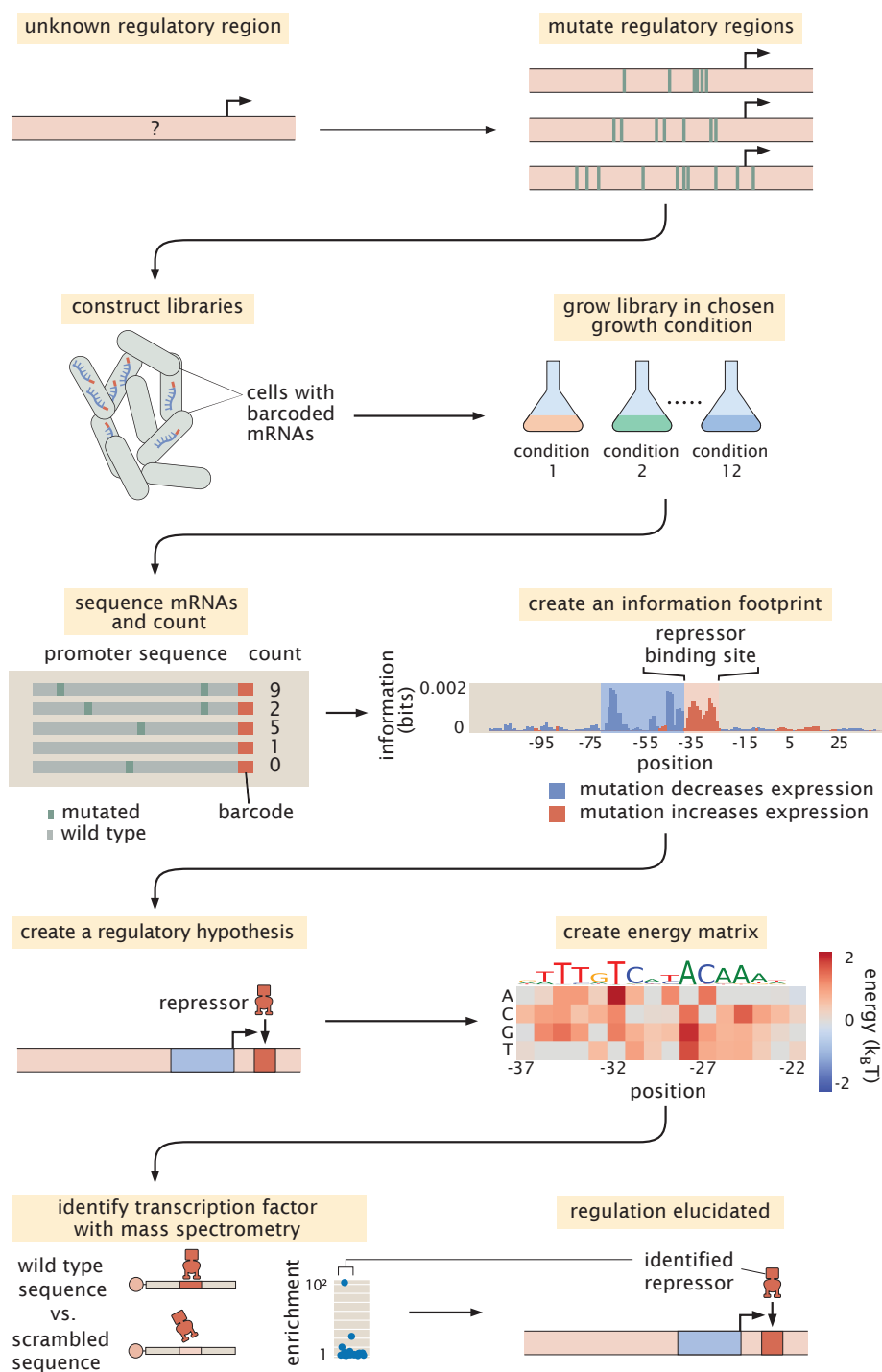


Figure 2.2: Schematic of the Reg-Seq procedure as used to recover a repressor binding site. (Continued on the following page.)

Figure 2.2: The process is as follows: After constructing a promoter library driving expression of a randomized barcode (an average of 5 barcodes for each promoter), RNA-Seq is conducted to determine the frequency of these mRNA barcodes across different growth conditions (list included in Appendix 2.6 Section “Growth conditions”). By computing the mutual information between DNA sequence and mRNA barcode counts for each base pair in the promoter region, an "information footprint" is constructed that yields a regulatory hypothesis for the putative binding sites (with the RNAP binding region highlighted in blue and the repressor binding site highlighted in red). Energy matrices, which describe the effect that any given mutation has on DNA binding energy, as well as sequence logos, are inferred for the putative transcription factor binding sites. Next, we identify which transcription factor preferentially binds to the putative binding site via DNA affinity chromatography followed by mass spectrometry. This procedure culminates in a coarse-grained, cartoon-level view of our regulatory hypothesis for how a given promoter is regulated.

from our promoter of interest based on the base-pair by base-pair sequence of its promoter. Using the mutual information between mRNA counts and sequences, we develop an information footprint that reveals the importance of different bases in the promoter region to the overall level of expression. We locate potential transcription factor binding regions by looking for clusters of base pairs that have a significant effect on gene expression. Further details on how potential binding sites are identified are found in the Methods Section “Automated putative binding site algorithm” and “Manual selection of binding sites”, while determination of the false positive and false negative rates of the method can be found in Appendix 2.7 Section “False positive and false negative rates”. Blue regions of the histogram shown in the information footprints of Figure 2.2 correspond to hypothesized activating sequences and red regions of the histogram correspond to hypothesized repressing sequences.

With the information footprint in hand, we can then determine energy matrices and sequence logos (described in the next section). Given putative binding sites, we use synthesized oligonucleotides that serve as fishing hooks to isolate the transcription factors that bind to those putative binding sites using DNA-affinity chromatography and mass spectrometry (Mittler, Butter, and Mann, 2009). Given all of this information, we can then formulate a schematized view of the newly discovered regulatory architecture of the previously uncharacterized promoter. For the case schematized in Figure 2.2, the experimental pipeline yields a complete picture of a simple repression architecture (i.e. a gene regulated by a single binding site for a repressor).

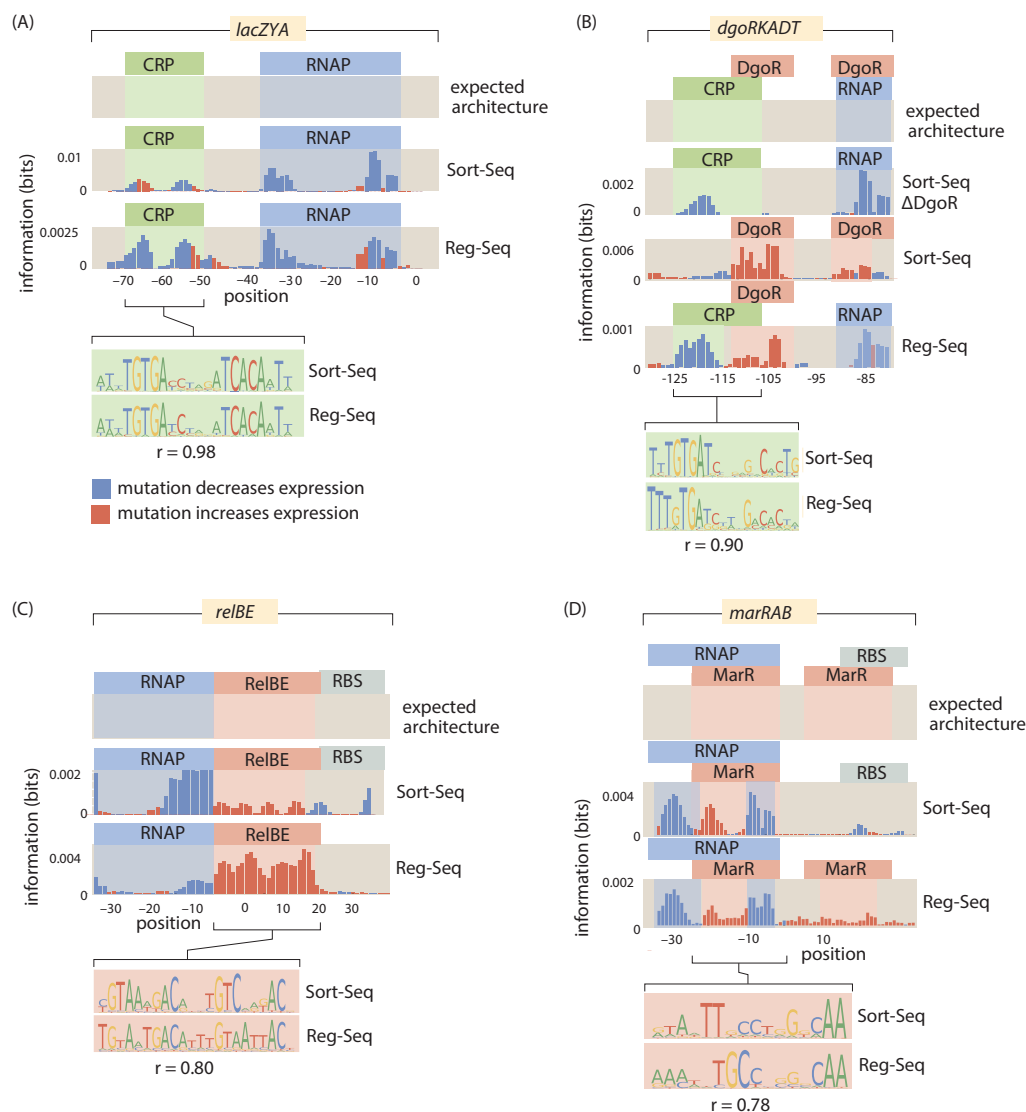


Figure 2.3: A summary of four direct comparisons of measurements from Sort-Seq and Reg-Seq. We show the identified regulatory regions as well as quantitative comparisons between inferred position weight matrices. (A) CRP binds upstream of RNAP in the *lacZYA* promoter. Despite the different measurement techniques for the two inferred position weight matrices, the CRP binding sites have a Pearson correlation coefficient of $r = 0.98$. (B) The *dgoRKADT* promoter is activated by CRP in the presence of galactonate and is repressed by DgoR. For Sort-Seq and Reg-Seq, type II activator binding sites can be identified based on the signals in the information footprint in the area indicated in green. Additionally the quantitative agreement between the CRP position weight matrices are strong, with $r = 0.9$. (C) The *relBE* promoter is repressed by RelBE as can be identified algorithmically in both Sort-Seq and Reg-Seq. The inferred logos for the two measurement methods have $r = 0.8$. (Continued on the following page)

Figure 2.3: (D) The *marRAB* promoter is repressed by MarR. The inferred energy matrices (data not shown) and sequence logos shown have $r = 0.78$. The right most MarR site overlaps with a ribosome binding site. The overlap has a stronger obscuring effect on the sequence specificity of the Sort-Seq measurement, which measures protein levels directly, than it does on the output of the Reg-Seq measurement.

Visual tools for data presentation

Throughout our investigation of the more than 100 genes explored in this study, we repeatedly relied on several key approaches to help make sense of the immense amount of data generated in these experiments. As these different approaches to viewing the results will appear repeatedly throughout the paper, here we familiarize the reader with five graphical representations referred to respectively as information footprints, energy matrices, sequence logos, mass spectrometry enrichment plots and regulatory cartoons, which taken together provide a quantitative description of previously uncharacterized promoters.

Information footprints: From our mutagenized libraries of promoter regions, we can build up a base-pair by base-pair graphical understanding of how the promoter sequence relates to level of gene expression in the form of the information footprint shown in Figure 2.2. In this plot, the bar above each base pair position represents how large of an effect mutations at this location have on the level of gene expression. Specifically, the quantity plotted is the mutual information I_b at base pair b between mutation of a base pair at that position and the level of expression. In mathematical terms, the mutual information measures how much the joint probability $p(m, \mu)$ differs from the product of the probabilities $p_{mut}(m)p_{expr}(\mu)$ which would be produced if mutation and gene expression level were independent. Formally, the mutual information between having a mutation at position b and level of expression is given by

$$I_b = \sum_{m=0}^1 \sum_{\mu=0}^1 p(m, \mu) \log_2 \left(\frac{p(m, \mu)}{p_{mut}(m)p_{expr}(\mu)} \right). \quad (2.1)$$

Note that both m and μ are binary variables that characterize the mutational state of the base of interest and the level of expression, respectively. Specifically, m can take the values

$$m = \begin{cases} 0, & \text{if } b \text{ is a mutated base} \\ 1, & \text{if } b \text{ is a wild-type base} \end{cases} \quad (2.2)$$

and μ can take on values

$$\mu = \begin{cases} 0, & \text{for sequencing reads from the DNA library} \\ 1, & \text{for sequencing reads originating from mRNA,} \end{cases} \quad (2.3)$$

where both m and μ are index variables that tell us whether the base has been mutated and if so, how likely that the read at that position will correspond to an mRNA, reflecting gene expression or a promoter, reflecting a member of the library. The higher the ratio of mRNA to DNA reads at a given base position, the higher the expression. $p_{mut}(m)$ in equation 2.1 refers to the probability that a given sequencing read will be from a mutated base. $p_{expr}(\mu)$ is a numeric value that gives the ratio of the number of DNA or mRNA sequencing counts to the total number of sequencing counts for each barcode.

Furthermore, we color the bars based on whether mutations at this location lowered gene expression on average (in blue, indicating an activating role) or increased gene expression (in red, indicating a repressing role). In this experiment, we targeted the regulatory regions based on a guess of where a transcription start site (TSS) will be, based on experimentally confirmed sites contained in RegulonDB (Santos-Zavaleta et al., 2019), a 5' RACE experiment (Mendoza-Vargas et al., 2009), or by targeting small intergenic regions so as to capture all likely regulatory regions. Further details on TSS selection can be found in the Methods Section “Library design and construction”. After completing the Reg-Seq experiment, we note that many of the presumed TSS sites are not in the locations assumed, the promoters have multiple active RNA polymerase (RNAP) sites and TSS, or the primary TSS shifts with growth condition. To simplify the data presentation, the '0' base pair in all information footprints is set to the originally assumed base pair for the primary TSS, rather than one of the TSS that was found in the experiment.

Energy matrices: Focusing on an individual putative transcription factor binding site as revealed in the information footprint, we are interested in a more fine-grained, quantitative understanding of how the underlying protein-DNA interaction is determined. An energy matrix displays this information using a heat map format, where each column is a position in the putative binding site and each row displays the effect on binding that results from mutating to that given nucleotide (given as a change in the DNA-transcription factor interaction energy upon mutation) (Berg and Hoppel, 1987; Stormo and Fields, 1998; Kinney et al., 2010). These energy matrices are scaled such that the wild type sequence is colored in white, mutations

that improve binding are shown in blue, and mutations that weaken binding are shown in red. These energy matrices encode a full quantitative picture for how we expect sequence to relate to binding for a given transcription factor, such that we can provide a prediction for the binding energy of every possible binding site sequence as

$$\text{binding energy} = \sum_{i=1}^N \varepsilon_i, \quad (2.4)$$

where the energy matrix is predicated on an assumption of a linear binding model in which each base within the binding site region contributes a specific value (ε_i for the i^{th} base in the sequence) to the total binding energy. Energy matrices are either given in A.U. (arbitrary units) or, for several cases where the gene has a simple repression or activation architecture with a single RNA polymerase (RNAP) site, are assigned $k_B T$ energy units following the procedure in Kinney et al., 2010 and validated on repression by *lac* repressor in Barnes et al., 2019. The details of how and when absolute units are determined can be found in Appendix 2.8 Section “Inference of scaling factors for energy matrices”.

Sequence logos: From an energy matrix, we can also represent a preferred transcription factor binding site with the use of the letters corresponding to the four possible nucleotides, as is often done with position weight matrices (Schneider and Stephens, 1990). In these sequence logos, the size of the letters corresponds to how strong the preference is for that given nucleotide at that given position, which can be directly computed from the energy matrix. This method of visualizing the information contained within the energy matrix is more easily digested and allows for quick comparison among various binding sites.

Mass spectrometry enrichment plots: As the final piece of our experimental pipeline, we wish to determine the identity of the transcription factor we suspect is binding to our putative binding site that is represented in the energy matrix and sequence logo. While the details of the DNA affinity chromatography and mass spectrometry can be found in the methods, the results of these experiments are displayed in enrichment plots such as is shown in the bottom panel of Figure 2.2. In these plots, the relative abundance of each protein bound to our site of interest is quantified relative to a scrambled control sequence. The putative transcription factor is the one we find to be highly enriched compared to all other DNA binding proteins.

Regulatory cartoons: The ultimate result of all these detailed base-pair-by-base-pair resolution experiments yields a cartoon model of how we think the given promoter

is being regulated. A complete set of cartoons for all the architectures considered in our study is presented later in Figure 2.4. While the cartoon serves as a convenient, visual way to summarize our results, it is important to remember that these cartoons are a shorthand representation of all the data in the four quantitative measures described above and are, further, backed by quantitative predictions of how we expect the system to behave when tested experimentally. Throughout this paper, we use consistent iconography to illustrate the regulatory architecture of promoters with activators and their binding sites in green, repressors in red, and RNAP in blue.

Newly discovered *E. coli* regulatory architectures

Elucidating individual promoters

With the tools outlined above, we are positioned to explore individual promoters, specifically those belonging to the part of the *E. coli* genome for which the function of the genes is unknown. Previously christened as the ‘y-ome’, Ghatak et al., 2019 surprisingly found that roughly 35% of the genes in *E. coli* lack experimental evidence of function. The situation is likely worse for other organisms. For many of the genes in the y-ome, we remain similarly ignorant of how those genes are regulated. Figures 2.4 and 2.5 provide several examples of genes which until now had unknown regulation. As shown in Figure 2.5, our study has found the first examples that we are aware of in the entire *E. coli* genome of a binding site for YciT. These examples are intended to show the outcome of the methods developed here and to serve as an invitation to browse the online resource (<https://www.rpgroup.caltech.edu/RegSeq/interactive>) where our full dataset is presented.

The ability to find binding sites for both widely acting regulators and transcription factors which may have only a few sites in the whole genome allows us to get an in-depth and quantitative view of any given promoter. As indicated in Figures 2.5(A) and (B), we were able to perform the relevant search and capture for the transcription factors that bind our putative binding sites. In both of these cases, we now hypothesize that these newly discovered binding site-transcription factor pairs exert their control through repression. The ability to extract the quantitative features of regulatory control through energy matrices means that we can take a nearly unstudied gene such as *ykgE*, which is regulated by an understudied transcription factor YieP, and quickly get to the point at which we can do quantitative modeling in the style that we and many others have performed on the *lac* operon (Vilar and Leibler,

Figure 2.4: All regulatory architectures uncovered in this study. (Continued on the following page.)

Figure 2.4: For each regulated promoter, activators and their binding sites are labeled in green, repressors and their binding sites are labeled in red, and RNAP binding sites are labeled in blue. All cartoons are displayed with the transcription direction to the right. Only one RNAP site is depicted per promoter. The transcription factor binding sites displayed have either been identified by the method described in the Section “Automated putative binding site algorithm” or have additional evidence for their presence as described in Table 2.2. Binding sites found for these promoters in the EcoCyc or RegulonDB databases are only depicted in these cartoons if the sites are within the 160 bp mutagenized region studied, and are detected by Reg-Seq.

2003; Vilar, Guet, and Leibler, 2003; Bintu et al., 2005; Kinney et al., 2010; Garcia and Phillips, 2011; Vilar and Saiz, 2013; Barnes et al., 2019; Phillips et al., 2019).

A panoply of promoter results

Figure 2.6 (and Tables 2.1 and 2.2) provides a summary of the discoveries made in the work done here using our next-generation Reg-Seq approach. The outcome of our study is a set of hypothesized regulatory architectures as characterized by a suite of binding sites for RNAP, repressors, and activators, as well as the extremely potent binding energy matrices. We do not assume, *a priori*, that a particular collection of such binding sites is AND, OR, or any other logic (Galstyan et al., 2019). Figure 2.6(A) provides a shorthand notation that conveniently characterizes the different kinds of regulatory architectures found in bacteria. In this (n_a, n_r) notation, n_a and n_r correspond to the number of recovered activator and repressor binding sites, respectively. In previous work (Rydenfelt et al., 2014b), we have explored the entirety of what is known about the regulatory genome of *E. coli*, revealing that the most common motif is the $(0, 0)$ constitutive architecture, though we hypothesized that this is not a statement about the facts of the *E. coli* genome, but rather a reflection of our collective regulatory ignorance in the sense that we suspect that with further investigation, many of these apparent constitutive architectures will be found to be regulated under the right environmental conditions. The two most common regulatory architectures that emerged from our previous database survey are the $(0, 1)$ and the $(1, 0)$ architectures, the simple repression motif and the simple activation motif, respectively. It is interesting to consider that the $(0, 1)$ architecture is in fact the repressor-operon model originally introduced in the early 1960’s by Jacob and Monod as the concept of gene regulation emerged (Jacob and Monod, 1961). Now we see retrospectively the far reaching importance of that architecture across the regulatory genome.

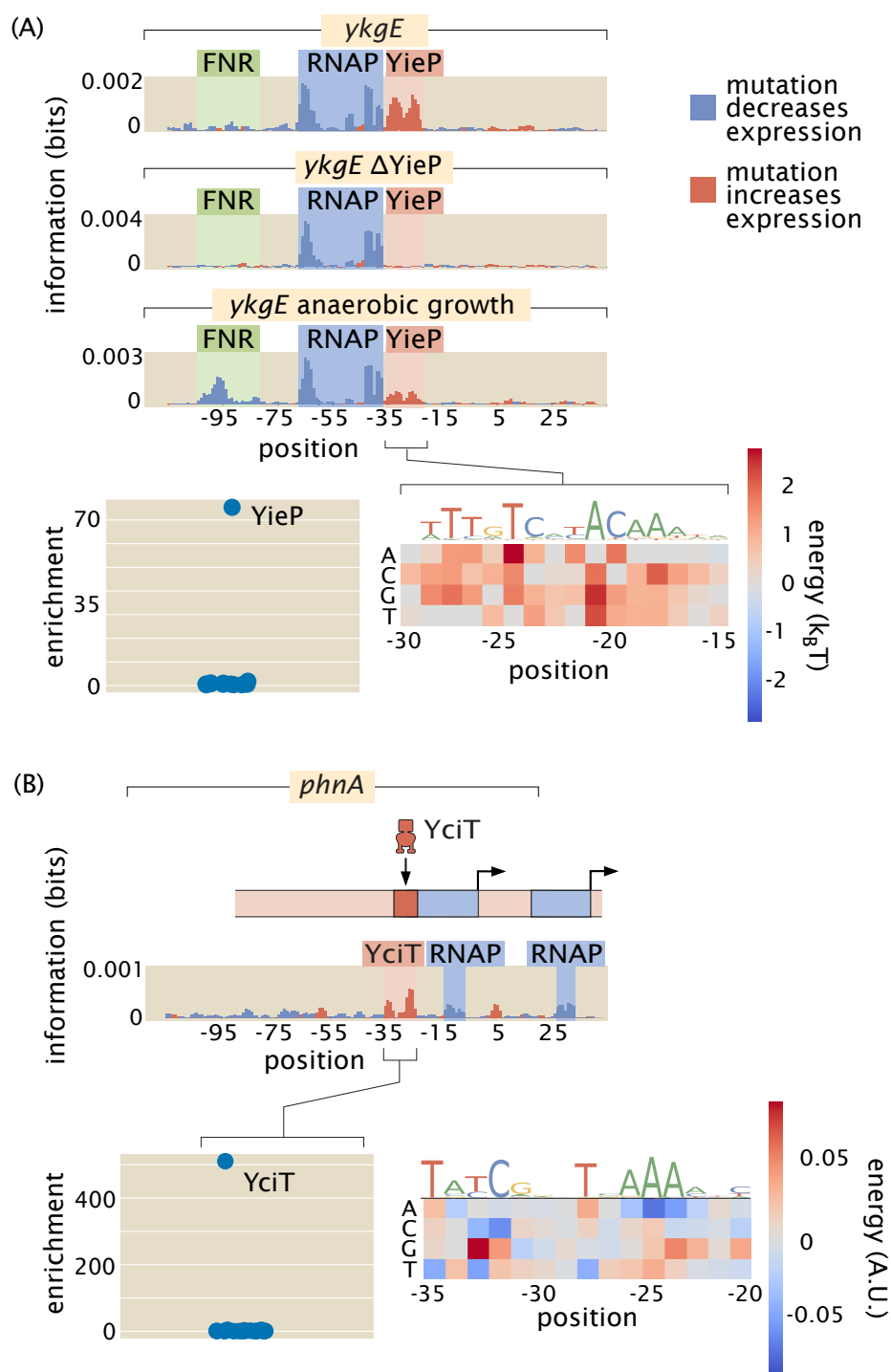


Figure 2.5: Examples of the insight gained by Reg-Seq in the context of promoters with no previously known regulatory information. (Continued on the following page.)

Figure 2.5: Activator binding regions are highlighted in green, repressor binding regions in red, and RNAP binding regions in blue. (A) From the information footprint of the *ykgE* promoter under different growth conditions, we can identify a repressor binding site downstream of the RNAP binding site. From the enrichment of proteins bound to the DNA sequence of the putative repressor as compared to a control sequence, we can identify YieP as the transcription factor bound to this site as it has a much higher enrichment ratio than any other protein. Lastly, the binding energy matrix for the repressor site along with corresponding sequence logo shows that the wild type sequence is the strongest possible binder and it displays an imperfect inverted repeat symmetry. (B) Illustration of a comparable dissection for the *phnA* promoter.

For the 113 genes we considered, Figure 2.6(B) summarizes the number of simple repression (0, 1) architectures discovered, the number of simple activation (1, 0) architectures discovered and so on. A comparison of the frequency of the different architectures found in our study to the frequencies of all the known architectures in the RegulonDB database is provided in Appendix 2.9 Figure 2.19. Tables 2.1 and 2.2 provide a more detailed view of our results. As seen in Table 2.1, of the 113 genes we considered, 34 of them revealed no signature of any transcription factor binding sites and they are labeled as (0, 0). The simple repression architecture (0, 1) was found 26 times, the simple activation architecture (1, 0) was found 11 times, and more complex architectures featuring multiple binding sites (e.g. (1, 1), (0, 2), (2, 0), etc.) were revealed as well. Further, for 18 of the genes that we label "inactive", Reg-Seq did not reveal a potential RNAP binding site. The lack of observable RNAP site could be because the proper growth condition to get high levels of expression was not used, or because the mutation window chosen for the gene does not capture a highly transcribing TSS.

The tables also include our set of 15 "gold standard" genes for which previous work has resulted in a knowledge (sometimes only partial) of their regulatory architectures. We find that our method recovers the regulatory elements of these gold standard cases fully in 11 out of 15 cases, and the majority of regulatory elements in 2 of the remaining cases. Overall the performance of Reg-Seq in these gold-standard cases (for more details see Appendix 2.7 Figure 2.14) builds confidence in the approach. Further, the failure modes inform us of the blind spots of Reg-Seq. For example, we find it challenging to observe weaker binding sites when multiple strong binding sites are also present such as in the *marRAB* operon. The *araC* case study shows that Reg-Seq does not perform well when many repressor sites regulate the promoter.

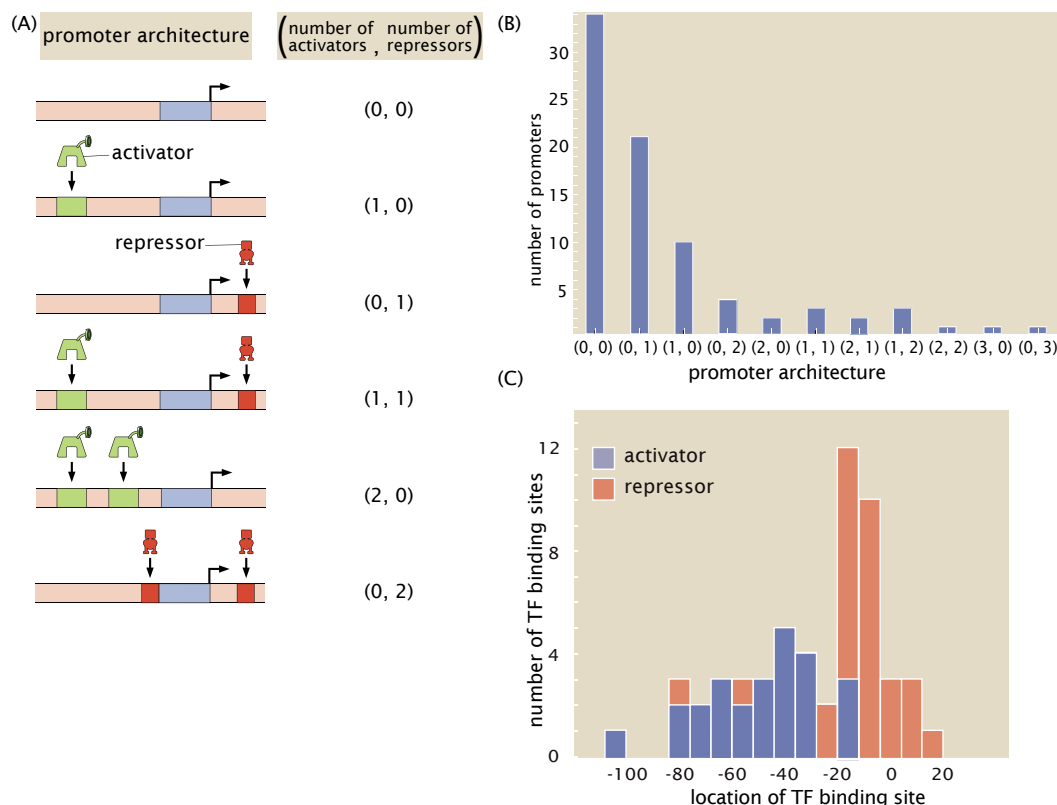


Figure 2.6: A summary of regulatory architectures discovered in this study. (A) The cartoons display a representative example of each type of architecture, along with the corresponding shorthand notation. (B) Counts of the different regulatory architectures discovered in this study. We exclude the "gold-standard" promoters (listed in Appendix 2.7 Table 2.4) unless new transcription factors are also discovered in the promoter. If, for example, one repressor was newly discovered and two activators were previously known, then the architecture is still counted as a (2,1) architecture. (C) Distribution of positions of binding sites discovered in this study for activators and repressors. Only newly discovered binding sites are included in this figure. The position of the transcription factor binding sites are calculated relative to the estimated TSS location, which is based on the location of the associated RNAP site.

Additionally the method will fail when there is no active TSS in the mutation window, as occurred in the case of *dicA*. Further details on the comparison to gold standard genes can be found in Appendix 2.7 Section "False positive and false negative rates".

We observe that the most common motif to emerge from our work (with the exception of constitutive expression) is the simple repression motif. Another relevant regulatory statistic is shown in Figure 2.6(C) where we see the distribution of bind-

Architecture	Total number of promoters	Number of promoters with at least one newly discovered binding site
All Architectures	113	48
(0,0)	34	0
(0,1)	26	21
(1,0)	11	10
(1,1)	4	3
(0,2)	4	4
(2,0)	3	2
(1,2)	4	3
(2,1)	2	2
(2,2)	1	1
(3,0)	3	1
(0,3)	2	1
(0,4)	1	0
inactive	18	0

Table 2.1: All promoters examined in this study, categorized according to type of regulatory architecture. Those promoters which have no recognizable RNAP site are labeled as inactive rather than constitutively expressed (0, 0).

ing site positions. Our own experience in the use of different quantitative modeling approaches to transcriptional regulation reveal that, for now, we remain largely ignorant of how to account for transcription factor binding site positions, and datasets like the one presented here will begin to provide data that can help us uncover how this parameter dictates gene expression. Indeed, with binding site positions and energy matrices in hand, we can systematically move these binding sites and explore the implications for the level of gene expression, providing a systematic tool to understand the role of binding-site position.

Uncovering the action of global regulators

One of the revealing case studies that demonstrates the broad reach of our approach for discovering regulatory architectures is offered by the insights we have gained into two widely acting regulators, GlpR (Figure 2.7) (Schweizer, Boos, and Larson, 1985) and FNR (Figure 2.8) (Körner, Sofia, and Zumft, 2003; Kargeti and Venkatesh, 2017). In both cases, we have expanded the array of promoters that they are now known to regulate. Further, these two case studies illustrate that even

for widely acting transcription factors, there is a large gap in regulatory knowledge and the approach advanced here has the power to discover new regulatory motifs. The newly discovered binding sites in Figure 2.7(A), with additional evidence for GlpR binding in Figure 2.7(B) and (C), more than double the number of operons known to be regulated by GlpR as reported in RegulonDB (Santos-Zavaleta et al., 2019). We found 5 newly regulated operons in our data set, even though we were not specifically targeting GlpR regulation. Although the number of example promoters across the genome that we considered is too small to make good estimates, finding 5 regulated operons out of approximately 100 examined operons supports the claim that GlpR widely regulates and many more of its sites would be found in a full search of the genome. The regulatory roles revealed in Figure 2.7(A) also reinforce the evidence that GlpR is a repressor.

For the GlpR-regulated operons newly discovered here, we found that this repressor binds strongly in the presence of glucose while all other growth conditions result in greatly diminished, but not entirely abolished, binding (Figure 2.7(A)). As there is no previously known direct molecular interaction between GlpR and glucose and the repression is reduced but not eliminated, the derepression in the absence of glucose is likely an indirect effect. As a potential mechanism of the indirect effect, *gpsA* is known to be activated by CRP (Seoh and Tai, 1999), and GpsA is involved in the synthesis of glycerol-3-phosphate (G3P), a known binding partner of GlpR which disables its repressive activity (Larson et al., 1987). Thus, in the presence of glucose, GpsA and consequently G3P will be found at low concentrations, ultimately allowing GlpR to fulfill its role as a repressor.

Architecture	Promoter	Newly discovered binding sites	Literature binding sites	Identified binding sites	Evidence
(0, 0)	<i>acuI</i>	0	0	0	
	<i>aegA</i>	0	0	0	
	<i>arcB</i>	0	0	0	
	<i>cra</i>	0	0	0	
	<i>dnaE</i>	0	0	0	
	<i>ecnB</i>	0	0	0	
	<i>fdoH</i>	0	0	0	
	<i>holC</i>	0	0	0	
	<i>hslU</i>	0	0	0	
	<i>htrB</i>	0	0	0	
	<i>minC</i>	0	0	0	
	<i>modE</i>	0	0	0	
	<i>ycgB</i>	0	0	0	
	<i>mscL</i>	0	0	0	
	<i>pitA</i>	0	0	0	
	<i>poxB</i>	0	0	0	
	<i>rlmA</i>	0	0	0	
	<i>rumB</i>	0	0	0	
	<i>sbcB</i>	0	0	0	
	<i>sdaB</i>	0	0	0	
	<i>tar</i>	0	0	0	
	<i>ybdG</i>	0	0	0	
	<i>ybiP</i>	0	0	0	
	<i>ybjT</i>	0	0	0	
	<i>yehT</i>	0	0	0	
	<i>yfhG</i>	0	0	0	
	<i>ygdH</i>	0	0	0	

continued on the following page

Architecture	Promoter	Newly discovered binding sites	Literature binding sites	Identified binding sites	Evidence
(0, 0)	<i>ygeR</i>	0	0	0	
	<i>yggW</i>	0	0	0	
	<i>ynaI</i>	0	0	0	
	<i>yqhC</i>	0	0	0	
	<i>zapB</i>	0	0	0	
	<i>zupT</i>	0	0	0	
	<i>amiC</i>	0	0	0	
(0, 1)	<i>araC</i>	0	1	0	
	<i>bdcR</i>	1	0	1	Known binding location (NsrR) (Partridge et al., 2009)
	<i>coaA</i>	1	0	0	
	<i>dicC</i>	0	1	0	
	<i>dinJ</i>	1	0	0	
	<i>ybeZ</i>	1	0	0	
	<i>idnK</i>	1	0	1	Mass-Spectrometry (YgbI)
	<i>leuABCD</i>	1	0	1	Mass-Spectrometry (YgbI)
	<i>mscM</i>	1	0	0	
	<i>yedK</i>	1	0	1	Mass-Spectrometry (TreR)
	<i>rapA</i>	1	0	1	Growth condition Knockout (GlpR), Bioinformatic (GlpR)
	<i>sdiA</i>	1	0	0	
	<i>tff-rpsB-tsf</i>	1	0	1	Growth condition Knockout (GlpR), Bioinformatic (GlpR), Knockout (GlpR)
	<i>thiM</i>	1	0	0	
	<i>tig</i>	1	0	1	Growth condition Knockout (GlpR), Bioinformatic (GlpR), Knockout (GlpR)
					<i>continued on the following page</i>

Architecture	Promoter	Newly discovered binding sites	Literature binding sites	Identified binding sites	Evidence
(0, 1)	<i>ybiO</i>	1	0	0	
	<i>ydjA</i>	1	0	0	
	<i>yedJ</i>	1	0	0	
	<i>phnA</i>	1	0	1	Mass-Spectrometry (YciT)
	<i>mutM</i>	1	0	0	
	<i>rhlE</i>	1	0	1	Growth condition Knockout (GlpR), Bioinformatic (GlpR), Mass-Spectrometry (GlpR)
	<i>uvrD</i>	1	0	1	Bioinformatic (LexA)
	<i>dusC</i>	1	0	0	
	<i>ftsK</i>	0	1	0	
	<i>znuA</i>	0	1	0	
	<i>znuCB</i>	0	1	0	
(1, 0)	<i>waaA-coaD</i>	1	0	0	
	<i>rcsF</i>	1	0	0	
	<i>groSL</i>	1	0	0	
	<i>mscS</i>	1	0	0	
	<i>thrLABC</i>	1	0	0	
	<i>yeiQ</i>	1	0	1	Growth condition Knockout (FNR), Bioinformatic (FNR)
	<i>ycbZ</i>	1	0	0	
	<i>ygiP</i>	1	0	0	
	<i>lac</i>	0	1	0	Bioinformatic (CRP)
	<i>yehS</i>	1	0	0	
	<i>yehU</i>	1	0	1	Growth condition Knockout (FNR), Bioinformatic (FNR)
(0, 2)	<i>pcm</i>	2	0	0	
	<i>yecE</i>	2	0	1	Mass-Spectrometry (HU)

continued on the following page

Architecture	Promoter	Newly discovered binding sites	Literature binding sites	Identified binding sites	Evidence
(0, 2)	<i>yjjJ</i>	2	0	1	Growth condition Knockout (MarA), Bioinformatic (MarA)
	<i>dcm</i>	2	0	1	Mass-Spectrometry (HNS)
(1, 1)	<i>arcA</i>	2	0	2	Growth condition Knockout (FNR), Bioinformatic (FNR), Mass-Spectrometry (FNR, CpxR)
	<i>dgoR</i>	0	2	0	Bioinformatic (CRP), Bioinformatic (DgoR)
	<i>ykgE</i>	2	0	2	Growth condition Knockout (FNR), Bioinformatic (FNR), Mass-Spectrometry (YieP), Knockout (YieP)
	<i>ymgG</i>	2	0	0	
(2, 0)	<i>asnA</i>	2	0	0	
	<i>fdhE</i>	2	0	2	Growth condition Knockout (FNR, ArcA), Bioinformatic (FNR, ArcA), Knockout (ArcA)
	<i>xylF</i>	0	2	0	
(1, 2)	<i>marR</i>	0	3	0	Mass-Spectrometry (MarR)
	<i>aphA</i>	3	0	2	Growth condition Knockout (FNR), Bioinformatic (FNR), Mass-Spectrometry (DeoR)
	<i>iap</i>	3	0	0	
(2, 1)	<i>ilvC</i>	3	0	1	Mass-Spectrometry (IlvY) (Rhee, Senear, and Hatfield, 1998)
	<i>maoP</i>	3	0	3	Growth condition Knockout (GlpR), Bioinformatic (GlpR), Knockout (PhoP, HdfR, GlpR)
	<i>rspA</i>	1	2	1	Mass-Spectrometry (DeoR)
					<i>continued on the following page</i>

Architecture	Promoter	Newly discovered binding sites	Literature binding sites	Identified binding sites	Evidence
(2, 2)	<i>ybjX</i>	4	0	4	Bioinformatic (2 PhoP sites), Mass-Spectrometry (HNS, StpA)
(3, 0)	<i>araAB</i>	0	3	0	
	<i>xylA</i>	0	3	0	
	<i>yicI</i>	3	0	0	
(0, 3)	<i>ompR</i>	0	3	0	
	<i>ybjL</i>	3	0	0	
(0, 4)	<i>relBE</i>	0	4	0	Mass-Spectrometry (RelBE)

Table 2.2: All genes investigated in this study categorized according to their regulatory architecture, given as (number of activators, number of repressors). The regulatory architectures as listed reflect **only** the binding sites that would be able to be recovered within our 160 bp constructs, but include both newly discovered and previously known binding sites. In those cases where binding sites that appear in RegulonDB or Ecocyc are omitted from this tally, the Section "Explanation of included binding sites" in Appendix 2.9 has the reasoning, for each relevant gene, why the binding sites are not shown. The table also lists the number of newly discovered binding sites, previously known binding sites, and number of identified transcription factors. The evidence used for the transcription factor identification is given in the final column. "Bioinformatic" evidence implies that discovered position weight matrices were compared to known transcription factor position weight matrices. The literature sites column contains only those sites that are both expected to be and are, in actuality, observed in the Reg-Seq data.

Prior to this study, there were 4 operons known to be regulated by GlpR, each with between 4 and 8 GlpR binding sites (Larson, Cantwell, and Loo-Bhattacharya, 1992; Zhao et al., 1994; Yang and Larson, 1996; Ye and Larson, 1988; Weissenborn, Wittekindt, and Larson, 1992), where the absence of glucose and the partial induction of GlpR was not enough to prompt a notable change in gene expression (Lin, 1976). These previously explored operons seemingly are regulated as part of an AND gate. *glpTQ*, *glpRABC*, *glpD*, and *glpFKX* have high gene expression when grown in growth media that does *not* contain glucose but does contain G3P (or glycerol, which leads to high concentrations of G3P). All other combinations of growth media, such as M9 glucose with G3P, or growth in LB without G3P, lead to low gene expression (Lin, 1976). In contrast, we have discovered operons whose regulation appears to be mediated by a single GlpR site per operon. With only a single site, GlpR functions as an indirect glucose sensor, as only the absence of glucose is needed to relieve repression by GlpR.

The second widely acting regulator our study revealed, FNR, has 151 binding sites already reported in RegulonDB and is well studied compared to most transcription factors (Gama-Castro et al., 2016). However, the newly discovered FNR sites displayed in Figure 2.8(A), with sequence logos of the respective sites displayed in Figure 2.8(B), demonstrate that even for well-understood transcription factors there is much still to be uncovered. Our information footprints are in agreement with previous studies suggesting that FNR acts as an activator. In the presence of O₂, dimeric FNR is converted to a monomeric form and its ability to bind DNA is greatly reduced (Myers et al., 2013). Only in low oxygen conditions did we observe a binding signature from FNR, and we show a representative example of the information footprint from one of 11 aerobic growth conditions in Figure 2.8(A).

We observe quantitatively how FNR affects the expression of *fdhE* both directly through transcription factor binding (Figure 2.9(B) and (C)) and indirectly through increased expression of ArcA (Figure 2.9(A), (B), (C), and (D)). Also, fully understanding even a single operon often requires investigating several regulatory regions as we have in the case of *fdoGHI-fdhE* by investigating the main promoter for the operon as well as the promoter upstream of *fdhE*. 36% of all multi-gene operons have at least one TSS which transcribes only a subset of the genes in the operon (Conway et al., 2014). Regulation within an operon is even more poorly studied than regulation in general. The main promoter for *fdoGHI-fdhE* has a repressor binding site, which demonstrates that there is regulatory control of the entire operon. How-

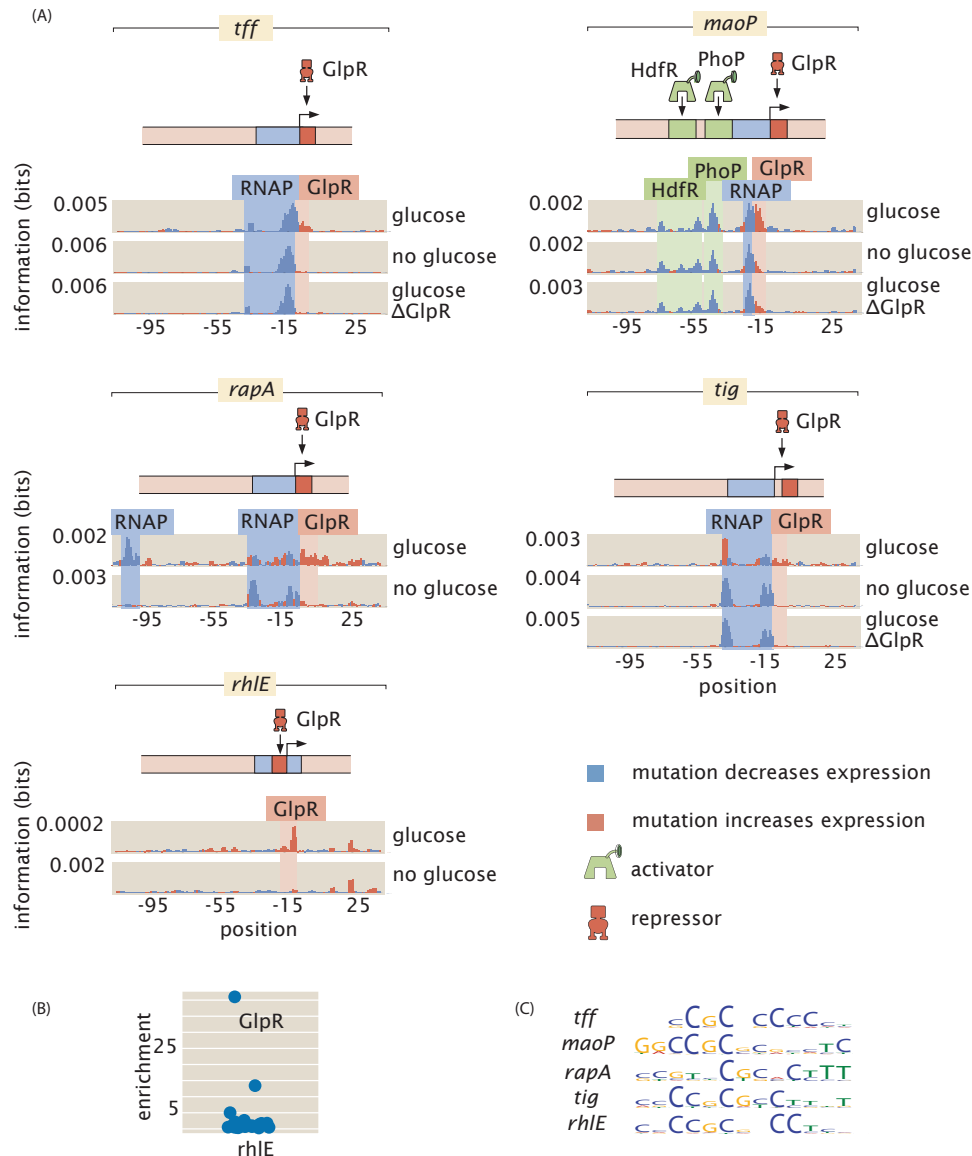


Figure 2.7: GlpR as a widely-acting regulator. (A) Information footprints for the promoters which we found to be regulated by GlpR, all of which were previously unknown. Activator binding regions are highlighted in green, repressor binding regions in red, and RNAP binding regions in blue. (B) GlpR was demonstrated to bind to *rhIE* by mass spectrometry. (C) Sequence logos for GlpR binding sites. Binding sites in the promoters of *tff*, *tig*, *maoP*, *rhIE*, and *rapA* have similar DNA binding preferences as seen in the sequence logos and each transcription factor binding site binds strongly only in the presence of glucose (As shown in (A)). These similarities suggest that the same transcription factor binds to each site. To test this hypothesis we knocked out GlpR and ran the Reg-Seq experiments for *tff*, *tig*, and *maoP*. In (A), we see that knocking out GlpR removes the binding signature of the transcription factor.

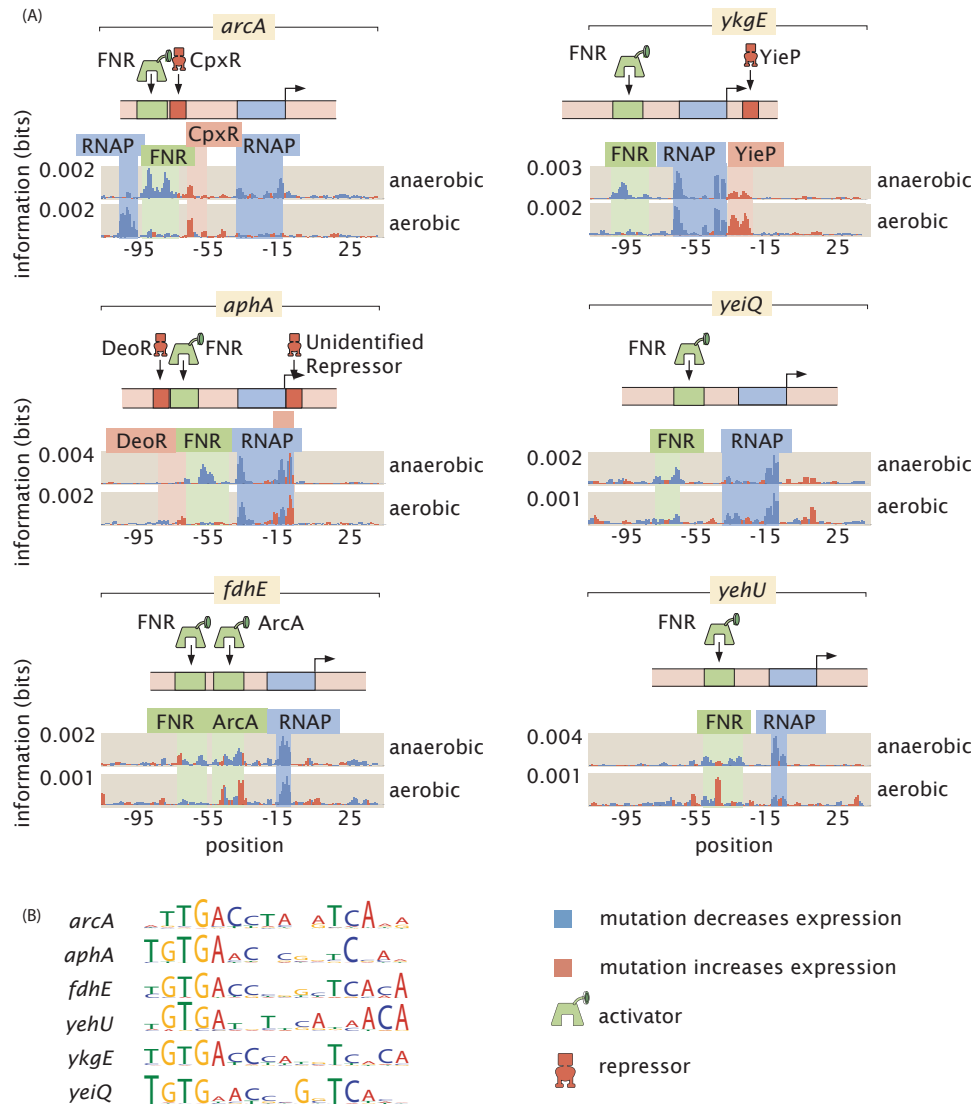


Figure 2.8: FNR as a global regulator. FNR is known to be upregulated in anaerobic growth, and here we found it to regulate a suite of six genes. In aerobic growth conditions the putative FNR sites are weakened. (A) Information footprints for the six regulated promoters. Activator binding regions are highlighted in green, repressor binding regions in red, and RNAP binding regions in blue. (B) Sequence logos for the FNR binding sites displayed in (A). The DNA binding preference of the six sites are shown to be similar from their sequence logos.

ever, we also see in Figure 2.9(B) that there is control at the promoter level, as *fdhE* is regulated by both ArcA and FNR and will therefore be upregulated in anaerobic conditions (Compan and Touati, 1994). The main TSS transcribes all four genes in the operon, while the secondary site shown in Figure 2.9(B) only transcribes *fdhE*,

and therefore anaerobic conditions will change the stoichiometry of the proteins produced by the operon. By investigating over a hundred promoter regions in this experiment it becomes feasible to target multiple promoters within an operon as we have done with *fdoGHI-fdhE*. We can then determine under what conditions an operon is internally regulated.

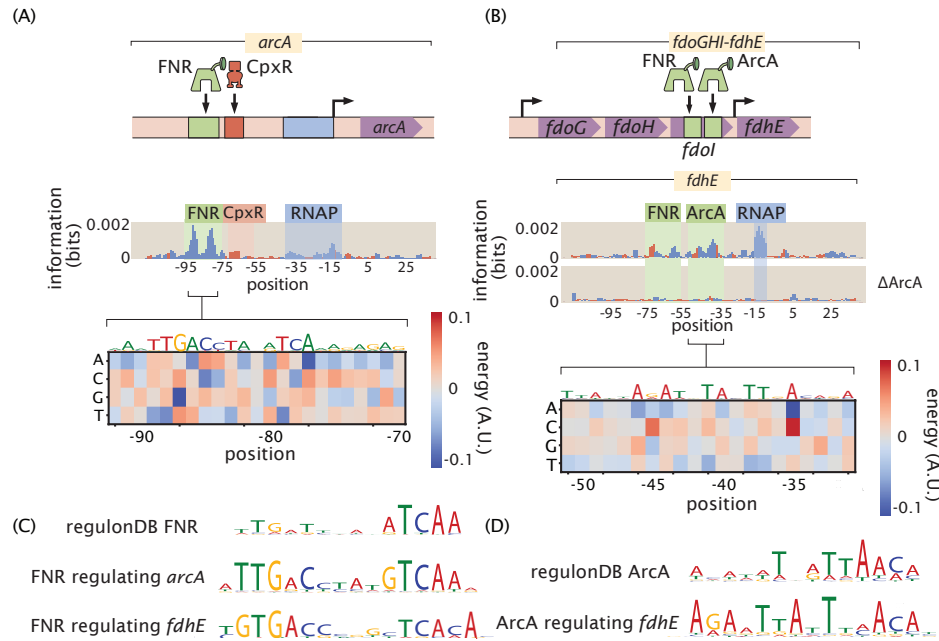


Figure 2.9: Inspection of a genetic circuit. (A) Here, the information footprint of the *arcA* promoter is displayed along with the energy matrix describing the discovered FNR binding site. (B) Intra-operon regulation of *fdhE* by both FNR and ArcA. The information footprint of *fdhE* is displayed. The discovered sites for FNR and ArcA are highlighted and the energy matrix for ArcA is displayed. A TOMTOM (Gupta et al., 2007) search of the binding motif found that ArcA was the most likely candidate for the transcription factor. The displayed information footprint from a knockout of ArcA demonstrates that the binding signature of the site, and its associated RNAP site, are no longer determinants of gene expression. (C) Sequence logos for FNR generated from both the sites cataloged in RegulonDB, as well as the discovered sites regulating *arcA* and *fdhE*. (D) Sequence logos for ArcA from sites contained in RegulonDB and the ArcA site regulating *fdhE*.

In summary

By examining the over 100 promoters considered here, grown under 12 growth conditions, we have a total of more than 1000 information footprints and data sets. In this age of big data, methods to explore and draw insights from that data are crucial.

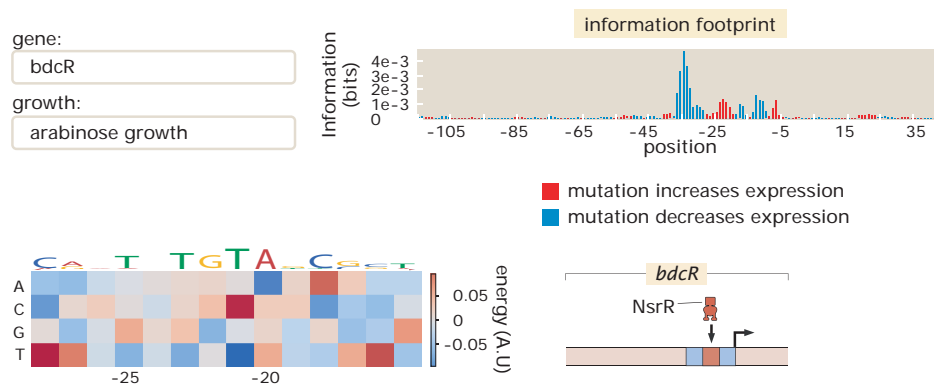


Figure 2.10: Representative view of the interactive figure that is available online. This interactive figure captures the entirety of our dataset. Each figure features a drop-down menu of genes and growth conditions. For each such gene and growth condition, there is a corresponding information footprint revealing putative binding sites, an energy matrix that shows the strength of binding of the relevant transcription factor to those binding sites and a cartoon that schematizes the newly-discovered regulatory architecture of that gene.

To that end, as introduced in Figure 2.10, we have developed an online resource (see <https://www.rpgroup.caltech.edu/RegSeq/interactive>) that makes it possible for anyone who is interested to view our data and draw their own biological conclusions. Information footprints for any combination of gene and growth condition are displayed via drop down menus. Each identified transcription factor binding site is marked, and energy matrices for all transcription factor binding sites are displayed. In addition, for each gene, we feature a simple cartoon-level schematic that captures our now current, best understanding of the regulatory architecture and resulting mechanism.

The interactive figure in question was invaluable in identifying transcription factors, such as GlpR, whose binding properties vary depending on growth condition. As sigma factor availability also varies greatly depending on growth condition, studying the interactive figure identified many of the secondary RNAP sites present. The interactive figure provides a valuable resource both to those who are interested in the regulation of a particular gene and those who wish to look for patterns in gene regulation across multiple genes or across different growth conditions.

2.4 Discussion

The study of gene regulation is one of the centerpieces of modern biology. As a result, it is surprising that in the genome era, our ignorance of the regulatory landscape in even the best-understood model organisms remains so vast. Despite understanding the regulation of transcription initiation in bacterial promoters (Browning and Busby, 2016), and how to tune their expression (Barnes et al., 2019), we lack an experimental framework to unravel understudied promoter architectures at scale. As such, in our view, one of the grand challenges of the genome era is the need to uncover the regulatory landscape for each and every organism with a known genome sequence. Given the ability to read and write DNA sequences at will, we are convinced that to make that reading of DNA sequence truly informative about biological function and to give that writing the full power and poetry of what Crick christened "the two great polymer languages", we need a full accounting of how the genes of a given organism are regulated and how environmental signals communicate with the transcription factors that mediate that regulation — the so-called "allosterome" problem (Lindsley and Rutter, 2006). The work presented here provides a general methodology for making progress on the former problem and also demonstrates that, by performing Reg-Seq in different growth conditions, we can make headway on the latter problem as well.

The advent of cheap DNA sequencing offers the promise of beginning to achieve this grand challenge in the form of MPRA as reviewed in Kinney and McCandlish, 2019. A particular implementation of such methods was christened Sort-Seq (Kinney et al., 2010) and was demonstrated in the context of well understood regulatory architectures. A second generation of the Sort-Seq method (Belliveau et al., 2018) established a full protocol for regulatory dissection through the use of DNA-affinity chromatography and mass spectrometry which made it possible to identify the transcription factors that bind the putative binding sites discovered by Sort-Seq. However, there were critical shortcomings in the method, not least of which was that it lacked the scalability to uncover the regulatory genome in a more multiplexed manner.

The work presented here builds on the foundations laid in previous studies by invoking RNA-Seq as a readout for the level of expression of the promoter mutant libraries needed to infer information footprints and their corresponding energy matrices and sequence logos. The original inference and hypothesis generation is followed by a combination of mass spectrometry, comparison of binding motifs, and gene knock-

outs to identify the transcription factors that bind those sites. The case studies described in the main text showcase the ability of the Reg-Seq method to deliver on the promise of beginning to uncover the regulatory genome systematically. The extensive online resources hint at a way of systematically reporting those insights in a way that can be used by the community at large to develop regulatory intuition for biological function and to design novel regulatory architectures using energy matrices.

However, several shortcomings remain in the approach introduced here. First, the current implementation of Reg-Seq is not fully automated for various aspects in the experimental pipeline; for example, manual examination of information footprints is used to generate testable regulatory hypotheses. As the method is scaled up further, this can limit throughput of the analysis. To address this for future work, we have created an automated methodology for identifying putative binding sites, which we describe in the methods section, that will simplify future scaled up efforts at identifying putative binding sites. All putative binding sites reported in this study either were identified through the automated methodology or have additional evidence for their presence such as mass spectrometry. In addition, these regulatory hypotheses can be converted into gene regulatory models using statistical physics (Buchler, Gerland, and Hwa, 2003; Bintu et al., 2005). However, here too, as the complexity of the regulatory architectures increases, it will be of great interest to use automated model generation as suggested in a recent biophysically-based neural network approach (Tareen and Kinney, 2019).

Another key challenge faced by the methods described here is that the mass spectrometry and the gene knockout confirmation aspects of the experimental pipeline remain low-throughput and, at times, inconclusive. Occasionally, we have found it challenging to observe weaker binding sites when multiple strong binding sites are also present. This was the case for the *marRAB* operon. To make our transcription factor identification methods more high-throughput, we have begun to explore a new generation of experiments such as *in vitro* binding assays that will make it possible to accomplish transcription factor identification in a multiplexed manner. Specifically, we are exploring multiplexed mass spectrometry measurements and multiplexed Reg-Seq on libraries of gene knockouts as ways to break the identification bottleneck. Transcription factor identification using Reg-Seq is also complicated by the growth conditions that we can test; for the 18 genes that we tested and labeled as "inactive" in this study, Reg-Seq did not reveal even an RNAP binding site, suggest-

ing that the proper growth condition to get high levels of expression was not used, or perhaps that the mutation window chosen for the gene does not capture a highly transcribing TSS. While information on the location of a TSS is available for 2500 of 2600 operons in *E. coli* (Santos-Zavaleta et al., 2019), this information does not guarantee those sites will have high transcription in the growth conditions studied. Similarly, many genes have multiple TSS that can be active under different growth conditions. In these cases we are limited both by the finite set of growth conditions we test as well as by the length of the mutation window, as it cannot always capture all TSS.

Another shortcoming of the current implementation of the method is that it misses regulatory action at a distance. Indeed, our laboratory has invested a significant effort in exploring such long-distance regulatory action in the form of DNA looping in bacteria (S. Johnson, Lindén, and Phillips, 2012; Han et al., 2009) and V(D)J recombination in jawed vertebrates (Lovely et al., 2015; Hirokawa et al., 2020). It is well known that transcriptional control through enhancers in eukaryotic regulation is central in contexts ranging from embryonic development to hematopoiesis (Melnikov et al., 2012). The current incarnation of the methods, as described here, have focused on contiguous regions in the vicinity of the transcription start site (within the 160 base pair mutagenized window). Clearly, to dissect the entire regulatory genome, these methods will have to be extended to non-contiguous regions of the genome.

Despite their limitations, the findings from this study provide a foundation for systematic, multiplexed regulatory dissections. We have developed a method to pass from complete regulatory ignorance to designable, regulatory architectures and we are hopeful that others will adopt these methods with the ambition of uncovering the regulatory architectures that preside over their organisms of interest.

2.5 Methods

Here, we provide an overview of the key methodological aspects of Reg-Seq. Extensive details of the methods used in this study can also be found on the GitHub Wiki associated with this work.

Library design and construction

We selected 113 TSS from the *E. coli* K12 genome for experiments. The promoter regions analyzed in this study were each 160 base pairs in length, a region that includes 45 base pairs downstream and 115 base pairs upstream of each TSS. The

general principles by which we selected each TSS were to first prioritize those TSS which have been extensively experimentally validated and catalogued in RegulonDB (Santos-Zavaleta et al., 2019) or EcoCyc (Keseler et al., 2017). Secondly, we selected those sites which had evidence of active transcription from RACE experiments (Mendoza-Vargas et al., 2009) and were listed in RegulonDB. If a TSS lacked both experimental evidence and active transcription as indicated by RACE experiments, we used the computationally predicted TSS as indicated on RegulonDB (Santos-Zavaleta et al., 2019) or EcoCyc (Keseler et al., 2017) and determined previously by (Huerta and Collado-Vides, 2003). If there were multiple TSS located upstream of the gene in question, we selected the TSS closest to the gene start, unless selecting one further upstream would allow for multiple TSS to be contained in the 160 base pair mutated region analyzed for each promoter. Not all TSS locations are known, and many genes have multiple TSS. The exact start sites used, as well as the reasoning behind our selection of each TSS, are listed in Supplementary File 1.

Promoter variants were synthesized on a microarray (TWIST Bioscience, San Francisco, CA). The sequences were designed computationally such that each base in the 160 base pair promoter region has a 10% probability of being mutated. For each promoter's oligonucleotide library, we ensured that the mutation rate as averaged across all sequences was kept between 9.5% and 10.5%, otherwise the library was regenerated. There are an average of 2200 unique promoter sequences per gene (for an analysis of how our results depend upon number of unique promoter sequences see Appendix 2.8 Figure 2.16). The library arrived lyophilized (76 pmol) and was resuspended in 100 μ L of TE (pH 8.0). 1 μ L of the resuspended oligonucleotide was amplified for 12 cycles with New England Biolabs Q5 High-Fidelity 2x Master Mix (NEB, Ipswich, MA) to increase the quantity of DNA in the library. Unless otherwise stated, all amplifications were performed using this polymerase mixture.

The PCR product was then run on a 2% TAE agarose gel, and approximately 200 base pair amplicons were extracted using a Zymoclean Gel DNA Recovery Kit (Zymo Research, Irvine, CA). To add a random 20-nucleotide barcode to each oligonucleotide, 1 ng of the purified DNA library was amplified for 10 PCR cycles using primers containing random 20-nucleotide DNA overhangs. All primer sequences can be found in Supplementary File 2. After cleaning this PCR product using a Zymo Clean and Concentrator Kit (Zymo Research, Irvine, CA), the library was cloned into the plasmid backbone of pJK14 (SC101 origin) (Kinney et al., 2010) using Gibson Assembly. An illustration of this plasmid is displayed in Ap-

pendix 2.6 Figure 2.12. Genetic constructs were electroporated into *E. coli* K-12 MG1655 (Blattner, 1997) and plated on LB plates with kanamycin. After 24 hours of growth on plates, libraries were scraped and inoculated into M9 media with 0.5% glucose in preparation for DNA sequencing.

All genetic barcodes were inserted 120 base pairs from the 5' end of the mRNA, containing 45 base pairs from the targeted regulatory region, 64 base pairs containing primer sites used in the construction of the plasmid, and 11 base pairs containing a three frame stop codon. Exact sequences of primers and spacer sequences for the constructs are listed in Supplementary File 2. Following each genetic barcode, there is an RBS, a GFP coding region, and a terminator.

Preparation of libraries for sequencing

To prepare cDNA libraries for sequencing, cells were grown to an optical density of 0.3 and RNA was stabilized using Qiagen RNA Protect (Qiagen, Hilden, Germany). Lysis was performed using lysozyme (Sigma Aldrich, Saint Louis, MO) and RNA isolated using the Qiagen RNA Mini Kit. Reverse transcription was preformed using Superscript IV (Invitrogen, Carlsbad, CA) with a specific primer for the labeled mRNA. qPCR was then performed in triplicate to check the level of DNA contamination. Any sample that had contaminating DNA at a level of 5% or more of the mRNA concentration was discarded. DNA libraries were prepared by growing cells to an optical density of 0.3 and isolating plasmid DNA with a spin miniprep kit (Qiagen, Hilden, Germany).

Sequencing

After preparing the barcoded libraries, we used next-generation sequencing (NGS) to map promoters to their respective barcodes. Sequencing libraries (both cDNA and DNA) had unindexed illumina flow cell adaptors attached via PCR, using primers that amplified a 221 base pair region that included the random barcode. We limited PCR cycles to exponential amplification, as determined by qPCR. Specifically, when we performed qPCR to check for DNA contamination, we also determined the number of cycles at which each sample reached exponential amplification, and then repeated the PCR reactions with the determined number of cycles to limit bias. After amplification, libraries were cleaned using a Zymo Clean and Concentrator kit and analyzed on an Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA). Samples were submitted to NGX Bio (NGX Bio, South Plainfield, NJ) for 150 base pair paired-end sequencing on a Hi-Seq 2500 (Illumina, San Diego, CA). We typically

acquired 250 million total reads for mapping of libraries. Further details of how we process the sequences can be found in Appendix 2.6 Section “Sequencing Analysis” and the GitHub Wiki associated with this work.

To quantify relative gene expression values for each promoter mutant in our library, we next grew cells expressing the DNA libraries in various growth conditions to an OD600 of 0.3. DNA and cDNA libraries were prepared in the same way as stated previously, and were sequenced at the Millard and Muriel Jacobs Genetics and Genomics Laboratory at Caltech on a HiSeq 2500 with a 100 base pair single read flow cell. An average of 5 unique 20 base pair barcodes per variant promoter was used for the purpose of counting transcripts. Specifically, for each promoter variant the number of sequences from the DNA library and the number of sequences produced from mRNA are determined.

Determination of energy matrices.

Energy matrices are used to represent the binding energy contribution for each nucleotide in a DNA sequence. We use relative gene expression values, as determined by counting genetic barcodes from NGS data for each mutated variant of a given regulatory sequence, and infer the energy contribution of each nucleotide by maximizing the mutual information between the rank-ordered binding strength predictions from the energy matrix and the gene expression data. We also perform this maximization using MCMC. Further discussion of how energy matrices are inferred can be found in Appendix 2.8 Section “Energy matrix inference” and on the GitHub Wiki that accompanies this study.

In each energy matrix plot, a red box indicates that a mutation to a nucleotide in that position decreases the energy of transcription factor binding, while a blue box indicates that a mutation at a given nucleotide position increases transcription factor binding energy. Energy matrices are typically given in arbitrary units, but the method by which we can assign absolute units in k_bT is covered in Appendix 2.8 Section “Inference of scaling factors for energy matrices”.

DNA affinity chromatography and mass spectrometry

Upon identifying a putative transcription factor binding site, we used DNA affinity chromatography, as performed in (Belliveau et al., 2018), to isolate and enrich for the transcription factor of interest. In brief, we order biotinylated oligos of our binding site of interest (Integrated DNA Technologies, Coralville, IA) along with a control, "scrambled" sequence, that we expect to have no specificity for

the given transcription factor. We tether these oligos to magnetic streptavidin beads (Dynabeads MyOne T1; ThermoFisher, Waltham, MA), and incubate them overnight with whole cell lysate grown in the presences of either heavy (with ^{15}N) or light (with ^{14}N) lysine for the experimental and control sequences, respectively. The next day, proteins are recovered by digesting the DNA with the PtsI restriction enzyme (New England Biolabs, Ipswich, MA), whose cut site was incorporated into all designed oligos.

Protein samples were then prepared for mass spectrometry by either in-gel or in-solution digestion using the Lys-C protease (Wako Chemicals, Osaka, Japan). Liquid chromatography coupled mass spectrometry (LC-MS) was performed as previously described by (Belliveau et al., 2018), and is further discussed in Appendix 2.8 Section “Processing of mass spectrometry experiments”. SILAC labeling was performed by growing cells (Δ LysA) in either heavy isotope form of lysine or its natural form.

It is also important to note that while we utilized the SILAC method to identify the transcription factor identities, our approach doesn’t require this specific technique. Specifically, our method only requires a way to contrast between the copy number of proteins bound to a target promoter in relation to a scrambled version of the promoter. In principle, one could use multiplexed proteomics based on isobaric mass tags (Pappireddi, Martin, and Wühr, 2019) to characterize up to 10 promoters in parallel. Isobaric tags are reagents used to covalently modify peptides by using the heavy-isotope distribution in the tag to encode different conditions. The most widely adopted methods for isobaric tagging are the isobaric tag for relative and absolute quantitation (iTRAQ) and the tandem mass tag (TMT). This multiplexed approach involves the fragmentation of peptide ions by colliding with an inert gas. The resulting ions are resolved in a second MS-MS scan (MS2).

Only a subset (13) of all transcription factor targets were identified by mass spectrometry due to limitations in scaling the technique to large numbers of targets. The transcription factors identified by this method are enriched more than any other DNA binding protein, with $p < 0.01$ using the outlier detection method as outlined by Cox and Mann, 2008, with corrections for multiple hypothesis testing using the method proposed by Benjamini and Hochberg, 1995. Details on data processing can be found in Appendix 2.8 Section “Processing of mass spectrometry experiments”. A detailed explanation of all experimental and computational steps can be found in the GitHub Wiki that accompanies this work.

Construction of knockout strains

Conducting DNA affinity chromatography followed by mass spectrometry on putative binding sites resulted in potential candidates for the transcription factors that bind to the target region. For some cases, to verify that a given transcription factor is, in fact, regulating a given promoter, we repeated the RNA sequencing experiments on strains in which the transcription factor of interest has been knocked out.

To construct the knockout strains, we ordered strains from the Keio collection (Yamamoto et al., 2009) from the Coli Genetic Stock Center. These knockouts were put in a MG1655 background via phage P1 transduction and verified with Sanger sequencing. To remove the kanamycin resistance that comes with the strains from the Keio collection, we transformed in the pCP20 plasmid (Datsenko and Wanner, 2000), induced FLP recombinase, and then selected for colonies that no longer grew on either kanamycin or ampicillin, verifying both loss of the chromosomally integrated kanamycin resistance and the pCP20 plasmid which confers ampicillin resistance. Finally, we transformed our desired promoter libraries into the constructed knockout strains, allowing us to perform the RNA sequencing in the same context as the original experiments.

Automated putative binding site algorithm

We introduce a systematized way of identifying the locations of binding sites to supplement manual curation (described in the Section “Manual selection of binding sites”). As illustrated in Figure 2.11, for a given information footprint, we average over 15 base pair "windows". We then determine which base pairs are part of a regulatory region by setting an information threshold of 2.5×10^{-4} bits. Threshold selection is described in Appendix 2.7 Section “False positive and false negative rates”. All base pair positions that pass the information threshold were then joined into regulatory regions. We consider "activator-like" (mutation decreases expression) and "repressor-like" (mutation increases expression) base pairs separately. This means that it is possible to have overlapping repressor and activator binding sites identified. We join any base pair positions within 4 base pairs of each other into single regulatory regions. We then find the edges of the region by trimming off any base pairs at the edge that are below the information threshold (even if the 15 base pair average is above the threshold). While we can often resolve overlapping or nearby repressors from activators, a limitation of this method of identification is that it cannot resolve two activators or two repressors that are very close to each other or overlapping.

To identify RNAP binding sites, we compare the sequence preference (through energy matrices and sequence logos) to experimentally validated examples of RNAP sites. We have examples of energy matrices for the σ^{70} RNAP site from Belliveau et al., 2018. For energy matrices of other σ factor binding sites, such as σ^{32} and σ^{28} , we use energy matrices generated from within the Reg-Seq experiment itself. For a σ^{32} binding site, for example, we used the example from the *hslU* gene. For a σ^{28} binding site, we used the energy matrix generated from the *dnaE* gene. We "scan" the example energy matrices across the mutated region. For each position in the region, we calculate the Pearson correlation coefficient between the example RNAP energy matrix and the inferred energy matrix at that position. We find RNAP binding site locations by thresholding the Pearson correlation coefficients at a value of 0.45. When performing manual curation of binding sites, we visually compare the sequence logos of the example RNAP binding sites to the sequence logos of putative binding sites. Further details of the method to create energy matrices and compare them to known motifs are given in Appendix 2.8 Section "Energy matrix inference" and Appendix 2.8 Section "TOMTOM motif comparison", respectively. Further, a detailed discussion of energy matrix construction is provided in the Sequencing Analysis GitHub Wiki page that accompanies this work.

Manual selection of binding sites

Similarly to the automated method of locating putative binding regions, we look for regions of high mutual information in the information footprints. While there was no hard cut-off for mutual information values during manual curation, we select clusters of base pairs that have a similar average information value (2.5×10^{-4} bits) to that described in the Section "Automated putative binding site algorithm".

During manual curation of binding sites, we also disqualify any binding sites where there are only 3 or fewer base pairs with high values in the mutual information footprint. The logic behind this decision is that individual bases with very high mutual information can potentially indicate that a putative binding site is only active when a certain mutation occurs. In turn, the binding site would not be active in wild-type conditions. To explain why this is, consider that a typical binding site mutation, at any given base pair, will significantly *weaken* the binding site of interest. Therefore, each of those mutated base pairs is said to have a "large effect" on expression. For a very poor binding site that is not active in the wild-type case, most mutations will further weaken a site which already will have only a minor

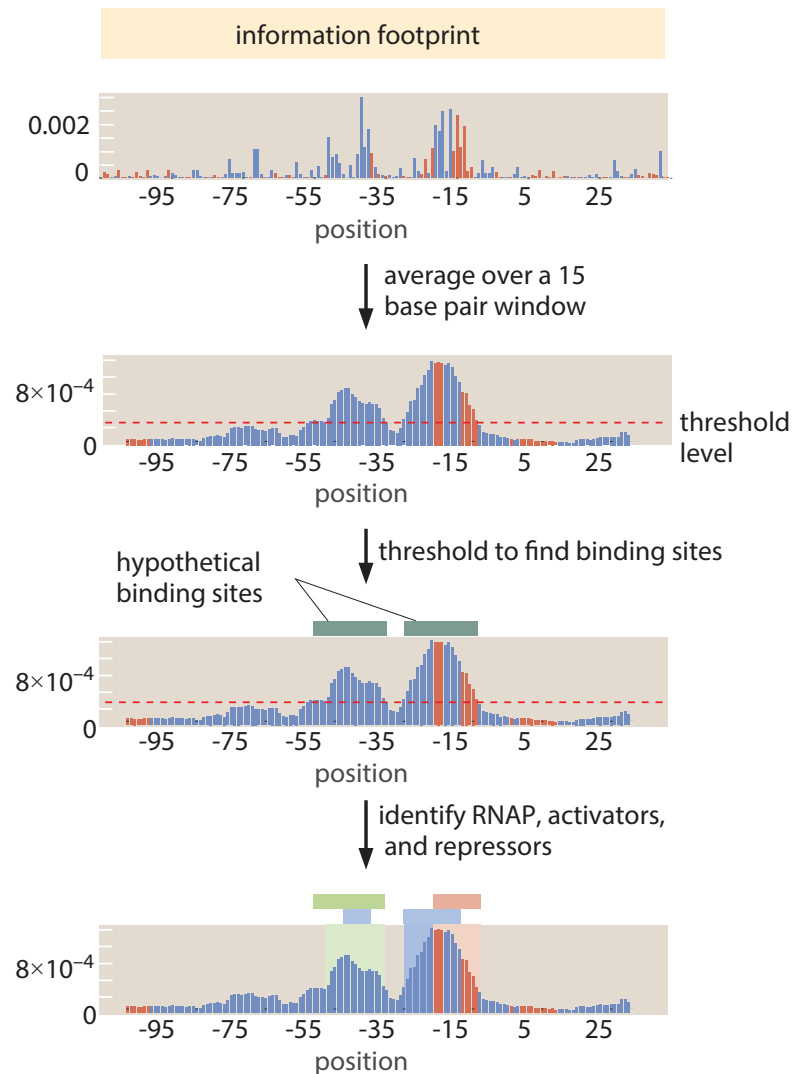


Figure 2.11: Procedure to identify binding site regions automatically. First, an information footprint is generated for the target region. Next, the information footprint is smoothed over a 15 base pair sliding window and a threshold of 2.5×10^{-4} bits is applied to identify regions of interest. RNAP binding sites are first identified (in blue), and the remainder of the regulatory regions are identified as repressor binding sites (if they tend to increase expression on mutation from wild type) or activator binding sites (if they tend to decrease expression upon mutation).

effect on gene expression. However, for a small number of base pairs, a mutation can occur that makes the DNA bind more tightly to the transcription factor, making it relevant for gene expression. Therefore, in the case of an extremely weak binding site that is not relevant in the wild type condition, there can still be a small number of highly informative bases. Initial hypothesis generation in Reg-Seq was done

manually. However, all those sites that are reported in Table 2.2 that do not have additional validation through mass spectrometry, gene knockouts, or bioinformatics appear in the set of putative binding sites generated by the method described in Section “Automated putative binding site algorithm”.

Code and Data Availability

An in-depth discussion of all experimental protocols and mathematical analysis used in this study can be found on the GitHub Wiki for this study (<https://github.com/RPGroup-PBoC/RegSeq/wiki>). All code used for processing data and plotting as well as the final processed data, plasmid sequences, and primer sequences can also be found on the GitHub repository (archived by Zenodo; <https://doi.org/10.5281/zenodo.3953312>). Energy matrices were generated using the MPAthic software (Ireland and Kinney, 2016). All raw sequencing data is available at the Sequence Read Archive (accession no. PRJNA599253 and PRJNA603368). All inferred information footprints and energy matrices can be found on the GitHub repository (archived by Zenodo; <https://doi.org/10.5281/zenodo.3953312>). All mass spectrometry raw data is available on the CaltechData repository (<https://doi.org/10.22002/d1.1336>).

2.6 Supplementary information: Extended details of experimental design

Choosing target genes

Genes in this study were chosen to cover several different categories. 29 genes had at least one transcription factor binding site listed in RegulonDB and were picked to validate our method under a number of conditions (15 with relevant high evidence sites). 37 were chosen because the work of Schmidt et al., 2016 demonstrated that gene expression changed significantly under different growth conditions. A handful of genes such as *minC*, *maoP*, or *fdhE* were chosen because we found either their physiological significance interesting, as in the case of *minC*, whose product is crucial for cell division and proper partitioning of the cell into two equal sized daughters in *E. coli* (Lutkenhaus, 2007). Alternatively, for some cases we found the gene regulatory question interesting, such as for the intra-operon regulation demonstrated by *fdhE*. The remainder of the genes were chosen because they had no regulatory information, often had minimal information about the function of the gene, and had an annotated transcription start site (TSS) in RegulonDB. A list of all genes chosen can be found in Supplementary File 1.

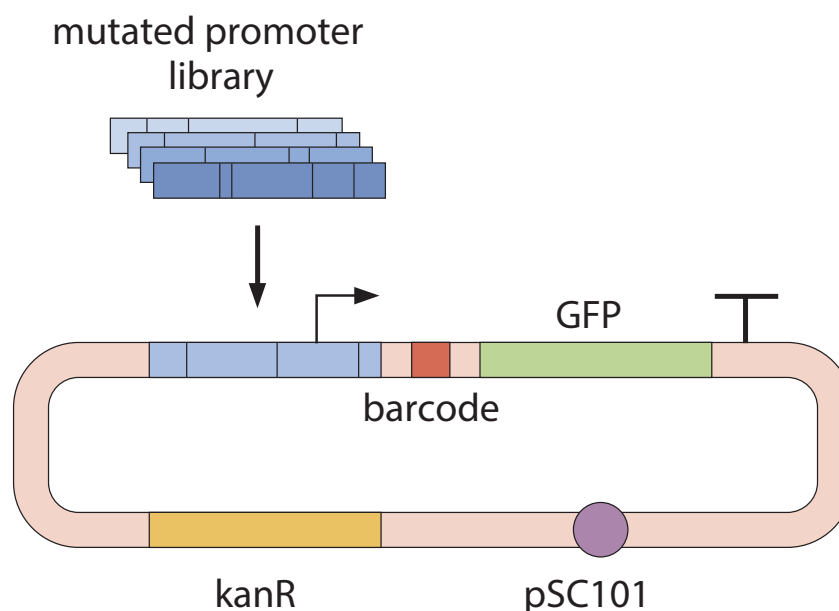


Figure 2.12: Schematic of the genetic construct used in this study. Mutated DNA libraries for each regulatory region were expressed from a pSC101 plasmid with kanamycin resistance (kanR). Each mutated sequence is 160 bp in length, which includes 45 bp downstream and 115 bp upstream of a given TSS. Each mutated sequence is flanked by primer binding sites to facilitate cloning. The genetic construct also contains a random barcode, a ribosome binding site (RBS), a GFP gene, and a terminator labeled with a large "T".

Sequencing Analysis

In this Appendix section, we provide further details associated with the analysis of next-generation sequencing (NGS) results, from both the "mapping" experiment, in which each unique barcode is "linked" to its corresponding mutated promoter region, and from the barcode sequencing experiments, in which the frequency of each barcode is counted and relative gene expression values determined. It is important to perform two sequencing experiments, in this manner, for a couple of reasons. Oligonucleotide libraries ordered from Twist Bioscience, which we use to construct promoter regions mutated at a 10% rate, are prone to random errors. This means that we do not fully know what is in the ordered library, and so it is necessary to sequence the full library and determine which mutations are present in each promoter region. The "mapping" phase of experiments also serves to connect each random, genetic barcode (which is added via PCR with primer overhangs) to

its corresponding, mutated promoter. By linking barcodes to promoters, we are able to build a "codex" that enables us to count genetic barcodes and, in turn, understand the relative gene expression values for each mutant promoter sequence.

For the "mapping" of genetic barcodes to their corresponding mutant promoter, we use paired-end sequencing, with 150 cycles for both Read 1 and Read 2, on a Hi-Seq 2500 machine. We acquired 250 million total reads for mapping of libraries.

In our analysis of FASTQ files, we removed any barcodes that were associated with a promoter variant which had insertions or deletions. Similarly, any genetic barcodes associated with multiple promoter variants were removed from the analysis, as were any sequences which appeared only once (barcodes must appear at least two times to be analyzed, as the appearance of a single, unique barcode sequence could be attributed to a sequencing error). The paired end reads from this sequencing step were assembled using the FLASH tool (Magoč and Salzberg, 2011). Any sequence with a PHRED score less than 20 was then removed using the FastX toolkit (Hannon, 2010). The specific commands used for this step of our analysis are listed on the GitHub Wiki associated with this work.

To analyze the "mapping" data and link each genetic barcode to its unique, mutagenized promoter region, we used a custom Python module, which can be found on the GitHub repository associated with this work. This module contains functions to check that sequences are the expected length, map unique barcodes to their corresponding promoter regions, and extract barcode sequences for subsequent sequencing experiments. We also provided a Jupyter notebook on the GitHub repository which provides a step-by-step walkthrough of the code used in processing sequencing data.

After mapping each barcode to its corresponding, mutated promoter region, we next "count" barcodes, both DNA and cDNA, to determine the relative gene expression values for each mutated promoter. For barcode counting experiments, only the region containing the random, 20 bp barcode was sequenced. For each growth condition, each promoter library yielded 20,000 to 500,000 usable sequencing reads. If the dataset for a gene in a given growth condition did not have at least 20,000 reads, it was not analyzed further, as we consistently found that, below this threshold, we reached a regime wherein the inference reliability of MCMC was reduced.

When preparing DNA and cDNA for NGS, we add a 4nt barcode, via PCR, to the library isolated from each growth condition. These 4nt barcodes are used during

data analysis both to map each library to its particular growth condition and to keep track of biological replicates, while the 20 bp barcodes can be used to identify each mutated promoter region. We performed all experiments with two biological replicates.

After collecting the FASTQ files, we perform quality filtering with FastX. We then perform barcode splitting with the FastX toolkit to separate each FASTQ file based on its growth condition, as well as separate the sequencing files based on whether they are derived from the DNA or cDNA library. Each experimental condition (both biological replicates, RNA vs. DNA, and growth conditions) receives a unique, 4nt barcode sequence, which enables us to identify where each library came from. Full details of our sequencing analysis methodologies, as well as all Python scripts, can be found on the GitHub repository associated with this work.

Growth conditions

The growth conditions used in this study were inspired by Schmidt et al., 2016, a study which observed changes in the *E. coli* proteome under growth conditions similar to the ones presented. The growth conditions utilized in this study are tabulated in Appendix 2.6 Table 2.3. The growth conditions explored here involved a range of environmental perturbations including altering the carbon source, inducing stress, or introducing trace metals. Unless otherwise noted in the caption of Appendix 2.6 Table 2.3, the cells were grown in the medium at 37 °C until reaching an OD of 0.3, at which point the cells were harvested and the RNA extracted. These growth conditions were chosen so as to span a wide range of growth rates, as well as to illuminate any carbon source specific regulators.

All knockout experiment were performed in M9 with glucose except for the knockouts for *arcA*, *hdfR*, and *phoP* which were grown in LB.

2.7 Supplementary information: Validating Reg-Seq against previous methods and results

The work presented here is effectively a third-generation of the use of Sort-Seq methods for the discovery of regulatory architecture. The primary difference between the present work and previous generations (Kinney et al., 2010; Belliveau et al., 2018) is the use of RNA-Seq rather than fluorescence and cell sorting as a readout of the level of expression of our promoter libraries. As such, there are many important questions to be asked about the comparison between the earlier methods and this work. We attack that question in several ways. First, as shown in

Growth conditions

M9 with glucose (0.5%)
 M9 with acetate (0.5%)
 M9 with arabinose (0.5%)
 M9 with xylose (0.5%) and arabinose (0.5%)
 M9 with succinate (0.5%)
 M9 with trehalose (0.5%)
 M9 with glucose (0.5%) and 5 mM sodium salicylate
 LB
 heat shock in M9 with glucose (0.5%)
 LB in low oxygen
 zinc, 5 mM ZnCl in M9 with glucose (0.5%)
 iron, 5 mM FeCL in M9 with glucose (0.5%)
 no cAMP in M9 with glucose (0.5%)

Table 2.3: All growth conditions used in the Reg-Seq study. Heat shocked cells were exposed to 42 °C for 5 minutes upon reaching OD 0.3 as this is known to induce transcription by σ^{32} (Arsène, Tomoyasu, and Bukau, 2000). Low oxygen growth cells were grown in a flask sealed with parafilm with minimal oxygen, although some was present as no anaerobic chamber was used. This level of oxygen stress was still sufficient to activate FNR binding, thus activating anaerobic metabolism. For cells grown with iron, upon reaching OD of 0.3 iron was added and cells were incubated for 10 minutes before harvesting RNA. Growth without cAMP was accomplished by the use of the JK10 strain (Kinney et al., 2010) which does not maintain its cAMP levels.

Figure 2.3, we have performed a head-to-head comparison of the two approaches to be described further in this section. Second, as shown in the next section, our list of candidate promoters included roughly 20% for which there is at least one experimentally validated transcription factor binding site. In these cases, we examined the extent to which our methods recover the known features of regulatory control about those promoters.

Comparison between Reg-Seq by RNA-Seq and fluorescent sorting

As the basis for comparing the results of the fluorescence-based Sort-Seq approach with our RNA-Seq-based approach, we use information footprints and position weight matrices as our metrics.

When making these comparisons between the two methods, we compare the values of a position weight matrix (PWM), often displayed as a sequence logo, generated

from the Sort-Seq and Reg-Seq methods. PWMs contain the probabilities that a given base will occur at a given position in the binding site. We calculate the Pearson correlation coefficient between the PWM values (represented as the height of the letters at each position) for the two methods. To compute the correlation coefficient, we use

$$r = \frac{\sum_{\alpha=1}^4 \sum_{i=1}^N (x_{i,\alpha} - \bar{x})(y_{i,\alpha} - \bar{y})}{\sqrt{\sum_{\alpha=1}^4 \sum_{i=1}^N (x_{i,\alpha} - \bar{x})^2} \sqrt{\sum_{\alpha=1}^4 \sum_{i=1}^N (y_{i,\alpha} - \bar{y})^2}}, \quad (2.5)$$

where $x_{i,\alpha}$ and $y_{i,\alpha}$ are the entries of the PWM of nucleotide α at position i obtained from Sort-Seq and Reg-Seq respectively, N is the total length of the binding site, and \bar{x} and \bar{y} are the means of $x_{i,\alpha}$ and $y_{i,\alpha}$, respectively. As an example, consider the following sequence logo from a Sort-Seq experiment,

position	A	C	G	T
1	0.01	0.01	0.03	0.95
2	0.04	0.83	0.06	0.07
3	0.70	0.17	0.11	0.02
4	0.86	0.01	0.10	0.03

and the same region resulting from a Reg-Seq experiment:

position	A	C	G	T
1	0.01	0.04	0.03	0.92
2	0.05	0.85	0.07	0.03
3	0.74	0.14	0.09	0.03
4	0.81	0.02	0.13	0.04

We see that for both sequence logos, the preferred nucleotides from position 1 through 4 are T-C-A-A, as indicated by the values in bold. Plugging in these values into equation 2.5, we get a Pearson correlation coefficient of $r = 0.997$, indicating substantial agreement between the Sort-Seq and Reg-Seq methods in this example. As a way to visualize similarity, for each position in the sequence logo we can plot the numerical value as resulting from the Sort-Seq experiment ($x_{i,\alpha}$) vs. the corresponding value obtained from the Reg-Seq experiment ($y_{i,\alpha}$). Perfect correspondence between the methods would result in all the data lying on the $x = y$ line (Appendix 2.7 Figure 2.13).

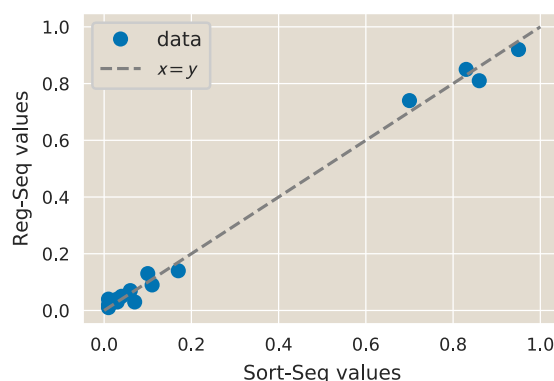


Figure 2.13: Mock data comparing Sort-Seq and Reg-Seq sequence logo values. These data have a Pearson correlation coefficient of $r = 0.997$. This high correlation is also indicated by the data deviating little from the $x = y$ line.

Figure 2.3 shows examples of this comparison for four distinct genes of interest. Figure 2.3(A) shows the results of the two methods for the *lacZYA* promoter with special reference to the CRP binding site. Both the information footprint and the position weight matrices (displayed with sequence logos) identify the same binding site.

Figure 2.3(B) provides a similar analysis for the *dgoRKADT* promoter where the position weight matrices for the CRP binding site from Reg-Seq and Sort-Seq have a correlation coefficient of $r = 0.90$. Figure 2.3(C) provides a quantitative dissection of the *relBE* promoter which is repressed by RelBE. Here we use both information footprints and expression shifts as a way to quantify the significance of mutations to different binding sites across the promoter. Finally, Figure 2.3(D) shows a comparison of the two methods for the *marRAB* promoter. The two approaches both identify a MarR binding site.

False positive and false negative rates.

We introduce a systematized way of identifying the locations of binding sites, as shown in Figure 2.11, that allows the false negative and false positive rate of binding site identification to be clearly assessed. For a given information footprint, we average over 15 base pair "windows". We then determine which base pairs are part of a regulatory region by setting an information threshold of 2.5×10^{-4} bits, which is explained below. All base pair positions that pass the information threshold are then joined into "regulatory regions", which we consider to be "activator-like"

(if a mutation decreases expression) or "repressor-like" (if a mutation increases expression). This means that it is possible to identify overlapping repressor and activator binding sites. We join any base pair positions within 4 base pairs of each other into a single regulatory region. We then find the edges of each binding site region by trimming off any base pairs at the edge that are below the information threshold (even if the 15 base pair average is above the threshold). A limitation of this method of identification is that it cannot resolve transcription factor binding sites that are very close to each other. The primary reasons for this is that putative binding sites will overlap after the smoothing step. While the method could be tuned to avoid treating nearby regions as the same site, many transcription factor binding sites will have sections of base pairs within their site where base identity has little to no effect on gene expression. Helix-turn-helix type transcription factors like CRP (whose binding site can be observed in Figure 2.3) are common examples of this phenomenon.

To determine which information threshold to use as a cutoff for a putative binding site, as displayed in Figure 2.11, we selected a training set of genes which included two of our "gold standard" genes with previously studied binding sites, DgoR (the upstream site from the *dgoR* promoter) and CRP (from the *araAB* promoter), two genes with only RNAP binding sites, including *hslU* (under heat shock) and *poxB*, and several genes that we classified as inactive, wherein no RNAP binding sites or other binding sites could be identified. These inactive genes included *hicB*, *mtgA*, *eco*, *hslU* (without heat shock), and *yncD*. The growth condition (heat shock) is specified for the *hslU* promoter as transcription occurs from a σ^{32} RNAP site, which will be inactive except during heat shock. We selected the threshold such that the RNAP sites and known binding sites were identified, while no binding sites were identified in the inactive regions.

We then determine a set of binding sites upon which to test this method and determine a false negative rate for the Reg-Seq experiment. In this set of binding sites, we include those sites which are "high-evidence" according to EcoCyc. Such "high evidence" binding sites have been validated experimentally with the binding of purified protein or through site mutation. Some "high-evidence" sites are excluded because they are not included within our 160 base pair, mutagenized sequence, or because they are not active in any of the growth conditions that we tested. Justifications for those binding sites which were not included are now listed in a new appendix; Appendix 2.9 Section "Explanation of included binding sites". A full list

gene	Transcription factor	Transcription factor type
<i>rspA</i>	CRP	activator
<i>rspA</i>	YdfH	repressor
<i>araAB</i>	AraC (2 sites)	activator
<i>znuCB</i>	Zur	repressor
<i>xylA</i>	CRP	activator
<i>xylA</i>	XylR (2 sites)	activator
<i>xylF</i>	XylR (2 sites)	activator
<i>dicC</i>	DicA	repressor
<i>relBE</i>	RelBE	repressor
<i>ftsK</i>	LexA	repressor
<i>znuA</i>	Zur	repressor
<i>lac</i>	CRP	activator
<i>marR</i>	Fis	activator
<i>marR</i>	MarA	activator
<i>marR</i>	MarR (2 sites)	repressor
<i>dgoR</i>	CRP	activator
<i>dgoR</i>	DgoR (right site)	repressor
<i>ompR</i>	IHF (3 sites)	repressor
<i>ompR</i>	CRP	repressor
<i>dicA</i>	DicA	repressor
<i>araC</i>	AraC (2 sites)	repressor
<i>araC</i>	AraC (2 sites)	activator
<i>araC</i>	CRP	activator
<i>araC</i>	XylR (2 sites)	repressor

Table 2.4: A suite of experimentally validated and high-evidence binding sites used to test our automated binding site finding algorithm. Specifically, this list of genes was used to test the false negative rate of our Reg-Seq method by examining what fraction of high-evidence sites were also identified with Reg-Seq.

of promoters and binding sites that *were* included in the set of genes used to validate our automated binding-site finding algorithms are also provided in Appendix 2.7 Table 2.4.

For each promoter contained in Appendix 2.7 Table 2.4, we used the automated procedure outlined above and in Figure 2.11 to identify the activator and repressor binding sites. A visual display of the expected binding sites, the information footprints for the promoters in Appendix 2.7 Table 2.4, and the discovered binding sites are all displayed in Appendix 2.7 Figure 2.14. To assess the false negative rate, we compare the identified regulatory regions to the known binding sites from

Appendix 2.7 Table 2.5. At this stage, we did not consider the identities of the binding sites; we merely consider their presence or absence. Inferred binding sites are declared to "match" the known binding site if the automated identification procedure classifies at least half of the base pairs reported in EcoCyc as belonging to a transcription factor binding site and correctly determines whether the binding site belongs to an activator or repressor.

We do not require exact matching of the edges of the binding sites for several reasons. One such reason is that, in some cases, the sequence of half of a binding site (for example, corresponding to one half of a helix-turn-helix binding motif) can contribute relatively little to gene expression, and so will not have high mutual information values in the corresponding information footprint for that binding site. While this may appear unintuitive for transcription factors where both sections of the binding site are bound by identical halves of a dimer, we see several examples of this in our Reg-Seq experiment results, including for CRP binding sites of the *rspA* promoter studied during our analysis of false negative rates. We can see in Appendix 2.7 Figure 2.14 that the downstream half of the binding site is not identified as important for gene expression. If we examine the wild type sequence of the *rspA* promoter, we also see that, for the upstream half of the sequence, the wild type matches the five most conserved bases of the consensus sequence (TGTGA) perfectly. The downstream half of the sequence, however, has 3 mismatches out of 5 bases. The downstream half of the binding site already binds to its target transcription factor poorly, so further mutations have little effect. While it is true that CRP binds to that sequence region, it is also true that CRP binds only extremely weakly to that section of the region. A similar effect can be seen in previous work from Belliveau et al., 2018, where a mutation in the downstream half of a CRP binding site in the *xyIE* promoter had more than a 10 fold greater effect on binding energy than mutation in the upstream half of the binding site. As such, we are lenient when evaluating the successes of our algorithm in this regard. Furthermore, the methods that have been used to determine the presence of "high evidence" binding sites in the past, such as ChIP-Seq, do not typically have base pair resolution with which to precisely determine the edges of binding sites (Skene and Henikoff, 2015).

Lastly, a known weakness of our algorithmic approach is that binding sites that are extremely close or overlapping cannot be distinguished from each other initially. For example, the XylR sites in the *xyIF* promoter are only separated by 3 bases according to RegulonDB. While the sites can be distinguished upon later investigation through

gene	Transcription factor	Was the region classified correctly?
<i>rspA</i>	CRP	Yes
<i>rspA</i>	YdfH	Yes
<i>araAB</i>	AraC (2 sites)	Yes
<i>znuCB</i>	Zur	Yes
<i>xylA</i>	CRP	Yes
<i>xylA</i>	XylR (2 sites)	Yes
<i>xylF</i>	XylR (2 sites)	Yes
<i>dicC</i>	DicA	Yes
<i>relBE</i>	RelBE	Yes
<i>ftsK</i>	LexA	Yes
<i>znuA</i>	Zur	Yes
<i>lac</i>	CRP	Yes
<i>marR</i>	Fis	No
<i>marR</i>	MarA	Yes
<i>marR</i>	MarR (2 sites)	Yes
<i>dgoR</i>	CRP	Yes
<i>dgoR</i>	DgoR (right site)	No
<i>ompR</i>	IHF (3 sites)	Yes
<i>ompR</i>	CRP	No
<i>dicA</i>	DicA	No
<i>araC</i>	AraC (4 sites)	1 site identified
<i>araC</i>	CRP	No
<i>araC</i>	XylR (2 sites)	No

Table 2.5: The results of the comparison between experimentally verified, high evidence binding sites and Reg-Seq binding sites. A visual illustration of the comparison can be found in Appendix 2.7 Figure 2.14.

gene knockouts, mass spectrometry, or motif comparison, our initial algorithm joins the sites into one large site. While this is a weakness of the algorithm, for our purposes it does not constitute a false negative, as the important regions for regulation are still discovered. All regions for all promoters that are classified as regulatory regions, their identities as activators, repressors, or RNAP binding sites, as well as their starting and ending base pairs, can be found in Supplementary File 3. Furthermore, we summarize the success and failures of the method at each binding site in Appendix 2.7 Table 2.5 below.

We see in Appendix 2.7 Table 2.5 that 11 of the 15 promoter regions included in Appendix 2.7 Table 2.4 have all transcription factor binding sites classified as

putative transcription factors, two have the majority of sites correctly classified, and two do not have any of their binding sites correctly classified as regulatory elements. We can see the information footprints used in the correct identifications in Appendix 2.7 Figure 2.14. We could alternatively consider that 23 out of 33 binding sites are correctly classified. However, we argue that the false negative rate should be considered on a per promoter basis, rather than on the basis of individual binding sites. The reason for this argument can be seen in the two "worst" cases of correct binding site identification, namely, for the *araC* and *dicA* promoters.

The *araC* promoter is repressed by multiple repressor binding sites in all growth conditions tested. *araC* only has high expression transiently after addition of arabinose (C. M. Johnson and Schleif, 1995), and while growth in arabinose is utilized in this experiment, RNA was not collected during the window of high expression. The case study shows that Reg-Seq does not perform well when many repressor sites regulate the promoter. Reg-Seq relies on mapping the effect on expression of mutating a particular site, and when many strong repressor sites are present, expression change will be minimal unless all repressor sites are weakened through mutation. Additionally, in this highly repressed case, the RNAP binding site we observe in the mutagenized region is not the documented RNAP site in RegulonDB, indicating that we are seeing transcription primarily from an alternative TSS. Different RNAP sites are often regulated differently, and in this case, the presence of an alternative and dominant RNAP binding site (in the repressed case), likely contributes to a failure to observe six of the seven binding sites in the *araC* promoter. Similarly, in the *dicA* promoter, we did not find an RNAP binding site in the studied region, which would make it very unlikely for any transcription factor binding sites to be identifiable.

In order to determine false positive rates, we test against promoters for which we are certain there are not additional, unannotated binding sites. Most known binding sites were not determined using a method like Reg-Seq, which looks for regulatory elements across an entire promoter region at base pair resolution. Rather, many efforts to pinpoint transcription factor binding site locations use assays like ChIP-Seq, which prioritizes looking for all binding sites of a given transcription factor across the entire genome. For those promoters studied with Reg-Seq, there are five promoters for which we have reason to believe that there are no undiscovered binding sites. There is evidence that the *zupT* promoter is constitutive (Grass et al., 2005), and the *marR*, *relBE*, *dgoR*, and *lacZYA* promoters have all been examined for binding sites at base pair resolution previously (in the Sort-Seq experiment

(Belliveau et al., 2018; Kinney et al., 2010)).

To evaluate false positive rates, we examine the putative activator and repressor binding sites as identified using our automated methodology (described previously), and compare any known binding sites to the known binding sites for the target promoters. We also classify any putative regulatory regions that are outside of known transcription factor binding sites as false positives. Similarly, any identified RNAP binding sites which were outside of the known RNAP binding locations were classified as false positives. In the *zupT* promoter, only the correctly placed RNAP site was identified. There were similarly no false positives identified in the *marR*, *relBE*, *dgoR*, or *lacZYA* promoters.

We additionally compare the energy matrices from putative regulatory regions to known binding site motifs. The known motifs are obtained either from RegulonDB or are generated from data from our prior Sort-Seq experiments (see (Belliveau et al., 2018)). We utilize the TOMTOM motif comparison software from (Gupta et al., 2007) to perform these comparisons. TOMTOM generates a p-value under the null hypothesis that the two compared motifs are drawn independently from the same underlying probability distribution. We test 95 motifs against each target motif that we are attempting to identify. The 95 resulting p-values (for each target) generated by TOMTOM are displayed in Appendix 2.7 Figure 2.15. A full discussion of TOMTOM can be found in Appendix 2.8 Section “TOMTOM motif comparison”. We only included those transcription factors that either have over 50 known binding sites in RegulonDB or have experimental measurements of binding site preference, such as in Sort-Seq (Belliveau et al., 2018). As such, we used TOMTOM on the XylR, CRP, MarA, MarR, and RelBE sites in Appendix 2.7 Table 2.5. We utilized a p-value cutoff of 0.05, corrected for multiple hypothesis testing. 95 motifs were tested against each target, and using the Bonferroni correction leads to a p-value cutoff of $\frac{0.05}{95} = 5 \times 10^{-4}$. In Appendix 2.7 Figure 2.15 we show that the correct transcription factor falls below the p-value threshold in all cases. For the CRP binding site in the *lacZYA* promoter, FNR also falls below the cutoff, but CRP has a calculated p-value that is ≈ 6 orders of magnitude lower, and so is clearly identified as the correct binding site. The results show that motif comparisons can be used reliably in those cases where we have high-quality energy matrices for comparison.



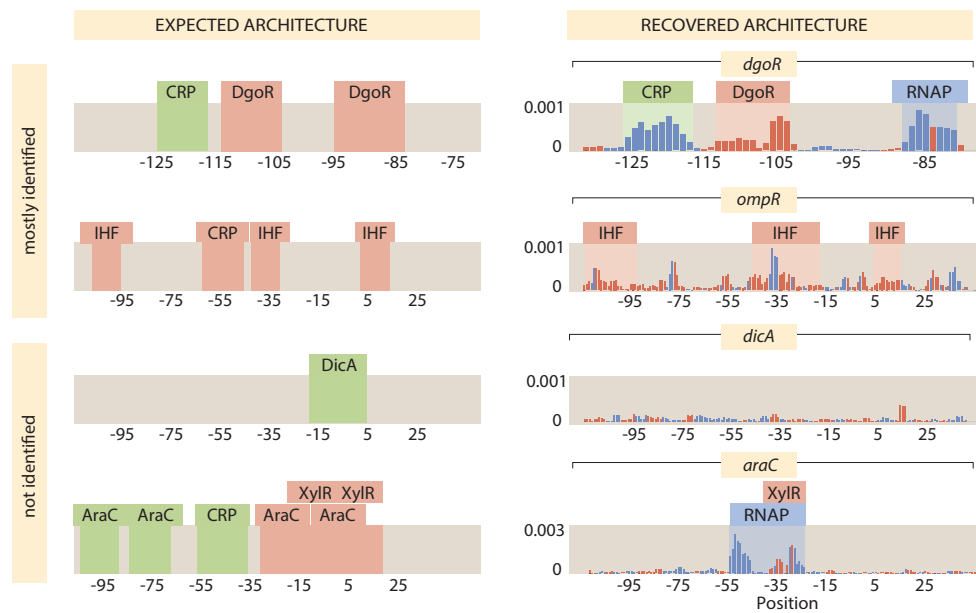


Figure 2.14: A visual comparison of the literature binding sites (left panel) and the extent of the binding sites discovered by our algorithmic approach (right panel). RNAP binding sites are also labeled in the right panel, but RNAP binding sites are not included in the false positive analysis.

2.8 Supplementary information: Extended details of analysis methods

Information footprints

We favor the use of information footprints as a tool for hypothesis generation to identify regions which may contain transcription factor binding sites. In general, a mutation within a transcription factor site is likely to weaken that site. We look for groups of positions where mutation away from wild type has a large effect on gene expression. Our datasets consist of nucleotide sequences, the number of times we sequenced a given, specific mutated promoter in the plasmid library, and the number of times we sequenced its corresponding mRNA. A simplified illustrative dataset on a hypothetical 4 nucleotide sequence is shown in Appendix 2.8 Table 2.6.

One strategy to measure the impact of a given mutation on expression is to take all sequences which have base b at position i and determine the number of mRNAs produced per read in the sequencing library. By comparing the values for different bases we can determine how large of an effect a mutation has on gene expression. For example in Appendix 2.8 Table 2.6, for the second position ($i = 2$) those sequences that contain the wild type base A ($b = A$) have 20 sequencing counts out of 50

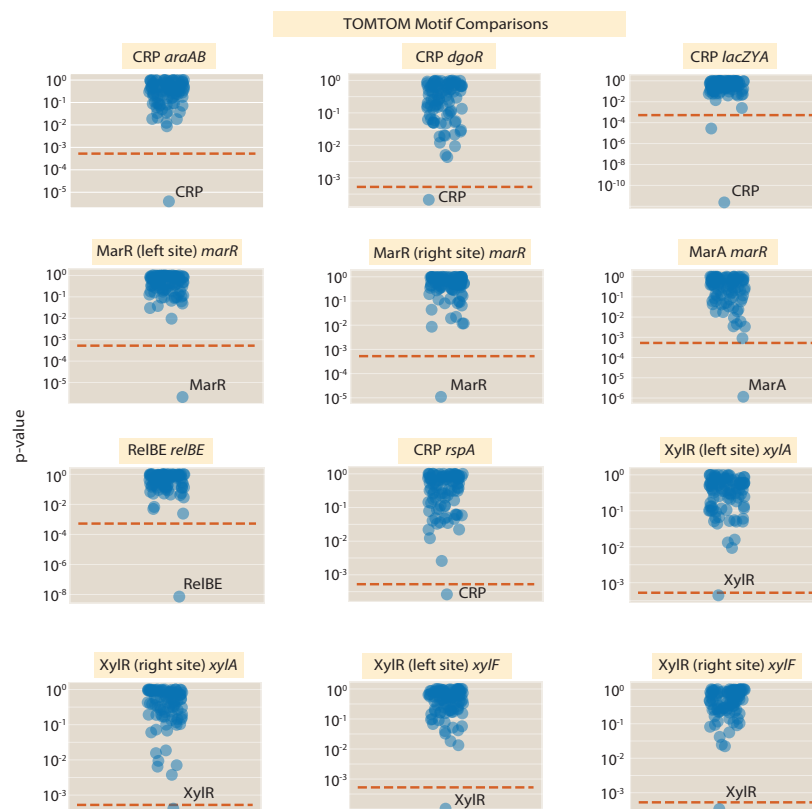


Figure 2.15: A visual display of the results of the TOMTOM motif comparison between the discovered binding sites and known sequence motifs from RegulonDB and our prior Sort-Seq experiment (Belliveau et al., 2018). Each dot in a given panel represents a comparison between the target position weight matrix (given in the figure title) and a position weight matrix for a given transcription factor. The p-value is calculated using the null hypothesis, that both motifs are drawn independently from the same underlying probability distribution. The red dotted line is displayed at a p-value of 5×10^{-4} . The line represents a p-value threshold of 0.05 that has been corrected for multiple hypothesis testing using the Bonferroni correction ($\frac{0.05}{95} = 5 \times 10^{-4}$).

Sequence	Library sequencing counts	mRNA counts
ACTA	5	23
ATTA	5	3
CCTG	11	11
TAGA	12	3
GTGC	2	0
CACA	8	7
AGGC	7	3

Table 2.6: Example dataset of 4 nucleotide sequences, and the corresponding counts from the plasmid library and mRNAs.

(23+3+11+3+0+7+3 = 50) from the DNA library and 10 sequencing counts from the 50 (5+5+11+12+2+8+7 = 50) mRNA reads. For all other sequences ($b = C, G$, or T), there are 30 sequencing counts from the DNA library and 40 sequencing counts from mRNA. A measure of the effect of mutation on expression would be to compare the ratios $\frac{\text{mRNA counts} / \text{total mRNA counts}}{\text{library counts} / \text{total library counts}}$ between mutated and wild type sequences. For the data in Appendix 2.8 Table 2.6, sequences with a wild type base at position 2 will have a ratio of purple(10/50)/(20/50) = 0.5 and sequences with a mutated base at position 2 will have a ratio of (40/50)/(30/50) \approx 1.3.

While directly comparing ratios is one way to measure the effect on gene expression, we use mutual information to quantify the effect of mutation, as Kinney et al., 2010 demonstrated could be done successfully. In Appendix 2.8 Table 2.6, the frequency of the nucleotide A in the DNA library at position 2 is 0.4, as 20 out of 50 sequencing counts have an A at position 2. Similarly, the other frequencies at position 2 are 0.32 for C, 0.14 for G and 0.14 for T. In the observed mRNA sequence counts, we find C at 34 of 50 total mRNA counts, which gives a frequency of 0.68, indicating that Cytosine is enriched in the mRNA transcripts compared to the DNA library. The frequencies for the other bases are 0.2 for A, 0.06 for T and 0.06 for G. Large enrichment of a base compared to others in mRNA sequencing counts occurs when base identity is important for gene expression.

We are classifying bases as either wild type ($m = 0$) or mutated ($m = 1$). A discussion of this assumption can be found at the end of this section. We compute mutual information at position i as

$$I_i = \sum_{m=0}^1 \sum_{\mu=0}^1 p(m, \mu) \log_2 \left(\frac{p(m, \mu)}{p_{mut}(m)p_{expr}(\mu)} \right), \quad (2.6)$$

where $p_{expr}(\mu)$ is the ratio of the number of DNA ($\mu = 0$) or mRNA ($\mu = 1$) sequencing counts to the total number of counts,

$$p_{expr}(\mu) = \begin{cases} \sum (\text{mRNA counts})/(\text{total counts}) & \text{if } \mu = 1 \\ \sum (\text{Library Sequencing counts})/(\text{total counts}), & \text{if } \mu = 0. \end{cases} \quad (2.7)$$

From the example data in Appendix 2.8 Table 2.6 we can calculate $p_{expr}(\mu)$. To do so, we sum up DNA counts and mRNA counts from all sequences and divide by the total number of counts ($50 + 50 = 100$) to obtain

$$p_{expr}(\mu) = \begin{cases} 0.5, & \text{if } \mu = 1 \\ 0.5, & \text{if } \mu = 0. \end{cases} \quad (2.8)$$

In addition, $p_{mut}(m)$ is the fraction of the total counts that either have a mutation ($m = 1$) at the given position or the fraction that have a wild type base ($m = 0$) at the position. p_{mut} has to be computed for each position individually. For position 1, the wild type base is an A, and we see that there are a total of 100 sequencing counts, of which 46 counts (DNA and mRNA combined) contain an A at position 1. Therefore $p(m)$ can be calculated for position 1 as

$$p_{mut}(m) = \begin{cases} 0.46, & \text{if } m = 0 \\ 0.54, & \text{if } m = 1. \end{cases} \quad (2.9)$$

Lastly, the joint distribution $p(m, \mu)$ is the probability that a given sequencing read in the dataset will have expression level μ and mutation status m . $p(m, \mu)$ is calculated by dividing the number of sequencing reads at the chosen position with mutation status m and expression status μ by the total number of sequencing reads. In the case of the example dataset in Appendix 2.8 Table 2.6 and for $m = 0$ and $\mu = 0$, we sum the sequencing reads that are wild type at position 1 and also are in the DNA library. As there are 17 sequences that fit the criteria out of 100 total sequences, $p(m = 0, \mu = 0) = 0.17$. The other values of $p(m, \mu)$ can be calculated to be

$$p(m, \mu) = \begin{cases} 0.17, & \text{if } m = 0 \text{ (wild type base) and } \mu = 0 \text{ (DNA)} \\ 0.21, & \text{if } m = 1 \text{ (mutated base) and } \mu = 1 \text{ (RNA)} \\ 0.33, & \text{if } m = 1 \text{ and } \mu = 0 \\ 0.29, & \text{if } m = 0 \text{ and } \mu = 1. \end{cases} \quad (2.10)$$

The marginal distributions p_{expr} and p_{mut} can be obtained by summing over one of the two variables, i.e.,

$$p_{expr}(\mu) = \sum_m p(m, \mu), \quad (2.11)$$

$$p_{mut}(m) = \sum_{\mu} p(m, \mu). \quad (2.12)$$

Plugging the values calculated above into equation (2.6) yields a mutual information value of 0.06 bits at position 1. The unit is bits because the mutual information is computed with a logarithm of base 2. Other bases can be chosen, however, that results in different units for the mutual information.

Mutual information is a measurement that quantifies how much the measurement of one of two variables reduces uncertainty of the other variable. For example, very low mutual information means that by knowing one variable one gains no information about the other variable, while on the other hand high mutual information means that by knowing one variable our knowledge about the others increases. At a position where base identity matters little for expression level, there would be little difference in the frequency distributions for the library and mRNA transcripts. The entropy of the distribution would decrease only by a small amount when considering the two types of sequencing reads separately.

We seek to determine the effect on gene expression of mutating a given base. However, if mutation rates at each position are not fully independent such that $p(m_i, m_{i'}) \neq p(m_i)p(m_{i'})$, then the information value calculated in equation (2.6) will also encode the effect of mutation at correlated positions. For instance, if position i is part of an activator binding site, mutating it will have a large effect on gene expression. If position i' is not within the activator site, then mutating position i' will have minimal true effect on gene expression. However, if mutations at the two bases are correlated, mutating position i' will make it more likely for i , and therefore the activator binding site, to be mutated. Knowledge that i' is mutated is predictive of overall expression, and so position i' will have high mutual information according to equation (2.6), even though that position has no regulatory function. In our experiment we designed sequences to be synthesized such that each position had a probability of mutation that was independent of mutation at any other position. However, due to errors in the oligonucleotide synthesis process, additional mutations

in the ordered sequences were introduced. Sequencing our DNA libraries reveals that mutation at a given base pair can make mutation at another base pair more likely by up to 10%, where neighboring base pairs are the most likely to have correlations between mutations. This is enough to cloud the signature of most transcription factors in an information footprint calculated using equation (2.6).

We need to determine values for $p_i(m|\mu)$ when mutations are independent, and to do this we need to fit these quantities from our data. We assert that

$$\langle C_{\text{mRNA}} \rangle \propto e^{-\beta E_{\text{eff}}} \quad (2.13)$$

is a reasonable approximation to make, which we will justify by considering a number of possible regulatory scenarios. $\langle C_{\text{mRNA}} \rangle$ is the average number of mRNAs produced and E_{eff} is an effective binding energy for the sequence that can be determined by summing contributions from each position in the sequence independently. There are many possible underlying regulatory architectures, and those that have been discovered with Reg-Seq are summarized in Table 2.1. While we will show that under reasonable assumptions this approach is useful for any of these regulatory architectures, let us first consider the simple case where there is only an RNAP site in the region under study. We can write down an expression for average gene expression per cell as

$$\langle C_{\text{mRNA}} \rangle \propto p_{\text{bound}} \propto \frac{\frac{P}{N_{\text{NS}}} e^{-\beta E_P}}{1 + \frac{P}{N_{\text{NS}}} e^{-\beta E_P}}, \quad (2.14)$$

where p_{bound} is the probability that the RNAP is bound to DNA and is known to be proportional to gene expression in *E. coli* (Ackers, A. D. Johnson, and Shea, 1982; Buchler, Gerland, and Hwa, 2003; Garcia and Phillips, 2011), E_P is the energy of RNAP binding, N_{NS} is the number of nonspecific DNA binding sites, and P is the number of RNAP. If RNAP binds weakly then $\frac{P}{N_{\text{NS}}} e^{-\beta E_P} \ll 1$, and we can simplify equation (2.14) to

$$\langle C_{\text{mRNA}} \rangle \propto e^{-\beta E_P}. \quad (2.15)$$

Using this relation, we can compute the ratio of average mRNA counts in wild type

$\langle C_{\text{mRNA}}^{\text{WT}_i} \rangle$ to average mRNA counts in a mutant $\langle C_{\text{mRNA}}^{\text{Mut}_i} \rangle$ as

$$\frac{\langle C_{\text{mRNA}}^{\text{WT}_i} \rangle}{\langle C_{\text{mRNA}}^{\text{Mut}_i} \rangle} = \frac{e^{-\beta E_{P_{\text{WT}_i}}}}{e^{-\beta E_{P_{\text{Mut}_i}}}}, \quad (2.16)$$

$$\frac{\langle C_{\text{mRNA}}^{\text{WT}_i} \rangle}{\langle C_{\text{mRNA}}^{\text{Mut}_i} \rangle} = e^{-\beta(E_{P_{\text{WT}_i}} - E_{P_{\text{Mut}_i}})}, \quad (2.17)$$

where $E_{P_{\text{WT}_i}}$ is the binding energy of RNAP to the wild type binding site and $E_{P_{\text{Mut}_i}}$ is the binding energy of RNAP to the mutant binding site. Using the assumption that each position contributes independently to the binding energy, we can simplify the differences in energies to $E_{P_{\text{WT}_i}} - E_{P_{\text{Mut}_i}} = \Delta E_{P_i}$. We can now calculate the probability of finding a specific base in the expressed sequences. If the probability of finding a wild type base at position i in the DNA library is $p_i(m=0|\mu=0)$, then

$$p_i(m=0|\mu=1) = \frac{p_i(m=0|\mu=0) \frac{\langle C_{\text{mRNA}}^{\text{WT}_i} \rangle}{\langle C_{\text{mRNA}}^{\text{Mut}_i} \rangle}}{p_i(m=1|\mu=0) + p_i(m=0|\mu=0) \frac{\langle C_{\text{mRNA}}^{\text{WT}_i} \rangle}{\langle C_{\text{mRNA}}^{\text{Mut}_i} \rangle}}, \quad (2.18)$$

$$p_i(m=0|\mu=1) = \frac{p_i(m=0|\mu=0)e^{-\beta\Delta E_{P_i}}}{p_i(m=1|\mu=0) + p_i(m=0|\mu=0)e^{-\beta\Delta E_{P_i}}}. \quad (2.19)$$

Under certain conditions, we can also infer a value for $p_i(m|\mu=1)$ using a linear model when there are any number of activator or repressor binding sites. We will demonstrate this in the case of a single activator and a single repressor, although a similar analysis can be done when there are greater numbers of transcription factors. Define $p = \frac{P}{N_{NS}}e^{-\beta E_P}$ and $a = \frac{A}{N_{NS}}e^{-\beta E_A}$ where A is the number of activators, and E_A is the binding energy of the activator. Also define $r = \frac{R}{N_{NS}}e^{-\beta E_R}$ where R is the number of repressors and E_R is the binding energy of the repressor. Then we can compute the average number of produced mRNA as

$$\langle C_{\text{mRNA}} \rangle \propto p_{\text{bound}} \propto \frac{p + pae^{-\beta\epsilon_{AP}}}{1 + a + p + r + pae^{-\beta\epsilon_{AP}}}, \quad (2.20)$$

where ϵ_{AP} is the interaction energy of activators and the RNAP. One assumption we make is that activators and RNAP bind weakly to their binding sites ($a \ll 1$ and

$p \ll 1$) but interact strongly ($pae^{-\beta\epsilon_{AP}} \gg p$). Under this assumption RNAP and associated activators are much more likely to bind DNA as a unit than separately. The binding energy measurements by Forcier et al., 2018 support this assumption in the case of CRP in the *lac* operon. The DNA-protein binding energy of CRP is measured to be $-3.18 k_B T$ and the interaction energy between CRP and RNAP is measured to be $\epsilon_{AP} = -6.56 k_B T$. The copy number of CRP is $A \approx 4000$ (Schmidt et al., 2016), the copy number of RNAP is $P \approx 2000$ in slowly growing cells (Bremer and Dennis, 1996), and the RNAP binding energy for the wild type *lac* promoter is $E_P \approx -5.2 k_B T$ (Brewster, Jones, and Phillips, 2012). As $N_{NS} \approx 4.6 \times 10^6$, the value of a can be calculated to be $a \approx \frac{4000}{4.6 \times 10^6} e^{3.18} \approx 0.02$. Similarly p can be calculated to be $p \approx \frac{2000}{4.6 \times 10^6} e^{5.2} \approx 0.08$. Lastly, we can calculate $pae^{-\beta\epsilon_{AP}} \approx pae^{6.56} \approx 1$. We can see that these numbers satisfy the assumptions $a \ll 1$, $p \ll 1$, and $pae^{-\epsilon_{AP}} \gg p$. We can simplify equation (2.20) to

$$\langle C_{\text{mRNA}} \rangle \propto p_{\text{bound}} \propto \frac{pae^{-\beta\epsilon_{AP}}}{1 + r + pae^{-\beta\epsilon_{AP}}}. \quad (2.21)$$

The last assumption we make is that repressors bind very strongly ($r \gg 1$ and $r \gg pae^{-\epsilon_{AP}}$). To justify this assumption we once again look to the *lac* operon. Wild type LacI copy number is $R \approx 10$ and the wild type binding energy for the O1 operator is $E_R \approx -16 k_B T$ (Garcia and Phillips, 2011). We can use these values to compute $r \approx \frac{10}{4.6 \times 10^6} e^{16} \approx 20$. We can simplify equation (2.21) to

$$\langle C_{\text{mRNA}} \rangle \propto \frac{pae^{-\beta\epsilon_{AP}}}{r} \quad (2.22)$$

$$\langle C_{\text{mRNA}} \rangle \propto e^{-\beta(-E_P - E_A + E_R)}, \quad (2.23)$$

As we typically assume that RNAP binding energy, activator binding energy, and repressor binding can all be represented as sums of contributions from their constituent bases, the combination of the energies can be written as a total effective energy E_{eff} which is a sum of independent contributions from all positions within the binding sites.

We fit the parameters for each base using Markov Chain Monte Carlo Method (MCMC). Two MCMC runs are conducted using randomly generated initial conditions. We require both chains to reach sufficiently similar distributions to prove the convergence of the chains or we repeat the runs. During the analysis we artificially

treat mutation rates at all positions as equal, as we do not wish for mutation rate to play a role in mutual information calculations. The information values are smoothed by averaging with neighboring values.

By only considering wild type or mutated energy contributions to the total effective binding energy rather than having separate values for energy contributions from all four base pairs, our methods will not be accurate in the case of calculating mutual information at locations with degenerate base pairs. However, the information footprints are intended to be hypothesis generation tools that can identify transcription factor binding sites. As such, the most important test for the assumption that we can approximate effective energy contributions from all 4 bases as contributions from only wild type or mutated bases is to assess whether the approximation has any effect on determining binding site locations. We re-ran the false positive and false negative assessments discussed in Appendix 2.7 Section “False positive and false negative rates”, but instead calculated the effective energy parameters for producing information footprints as a sum of contributions from all four bases. We find that the literature binding sites that were properly identified, as summarized in Appendix 2.7 Table 2.5, are identically identified. Specifically, any site which was identified using the previous method is still identified and any site that failed to be identified is still not observed. Similarly, when we only fit effective energy parameters for mutated or wild type bases there are no false positives identified in the promoters for *marR*, *relBE*, *dgoR*, *zupT*, or *lacZYA*. There are also no false positives when repeating the procedure while considering all 4 bases in the effective energy fits, implying that the simplification to only considering mutated or wild type bases does not have an effect on our ability to identify binding sites.

Processing of mass spectrometry experiments

Mass spectrometry results were processed using MaxQuant (Cox and Mann, 2008; Cox et al., 2009). Spectra were searched against the UniProt *E. coli* K-12 database as well as a contaminant database (256 sequences). LysC was specified as the digestion enzyme. Proteins were considered if they were known to be transcription factors, or were predicted to bind DNA (using gene ontology term GO:0003677, for DNA-binding in BioCyc).

Uncertainty due to number of independent sequences

1400 promoter variants were ordered from TWIST Bioscience for each promoter studied. Due to errors in oligonucleotide synthesis, additional mutations are ran-

domly introduced into the ordered oligos. We have found that, as a result of these random, additional errors, the final number of variants received was an average of 2200 per promoter.

To demonstrate that our results are not strongly dependent on the number of sequences in each promoter library, and also to assess how a reduction in the number of sequences per promoter library could facilitate larger scale experiments in the future, we generated examples of smaller data sets by computationally sub-sampling the Reg-Seq experimental data from 7 mutated promoter libraries; (*maoP*, *hslU*, *rpsA*, *leuABCD*, *aphA*, *araC*, and *tig*). These promoters are representative of a large cross section of the variety of regulation we see in our study, as they include promoters with constitutive expression (*hslU*), simple repression (*leuABCD*, *tig*), simple activation (*aphA*), as well as more complicated regulatory architectures (*maoP*, *rpsA*, *araC*). Each sub-sampling was done three times, and we then use the Pearson correlation coefficient (Appendix 2.7 Section “Comparison between Reg-Seq by RNA-Seq and fluorescent sorting”) as a comparison metric between the inference based on the full data set and the computationally sub-sampled data sets.

Based on our analysis, the results of which are displayed in Appendix 2.8 Figure 2.16, we find that there is only a small effect on the resulting sequence logo until the library has been reduced to approximately 500 promoter variants. We could, therefore, reasonably lower the resolution of the experiment to approximately 1000 or fewer unique sequences before large deviations in the inference are experienced. Decreasing the number of unique sequences can give modest boosts to the number of genes that can be studied, but will not be able to give order of magnitude increases in the number of genes that can be explored.

Effect on calculated energy matrices when transcription factor copy number \approx plasmid copy number

Throughout this study, we utilize plasmids to express GFP from mutated promoters, and then use the ratio of mRNA/DNA, based on sequencing results, to handle the effect of variability in plasmid copy number between cells. It is necessary, however, to consider the situation wherein the plasmid copy number is of a similar magnitude to the transcription factor copy number, and whether this can impact the calculated energy matrices and binding energies. The genetic expression levels are determined not only by the binding energy, but also by the transcription factor availability, and so it is necessary to consider whether, for those cases where transcription factor

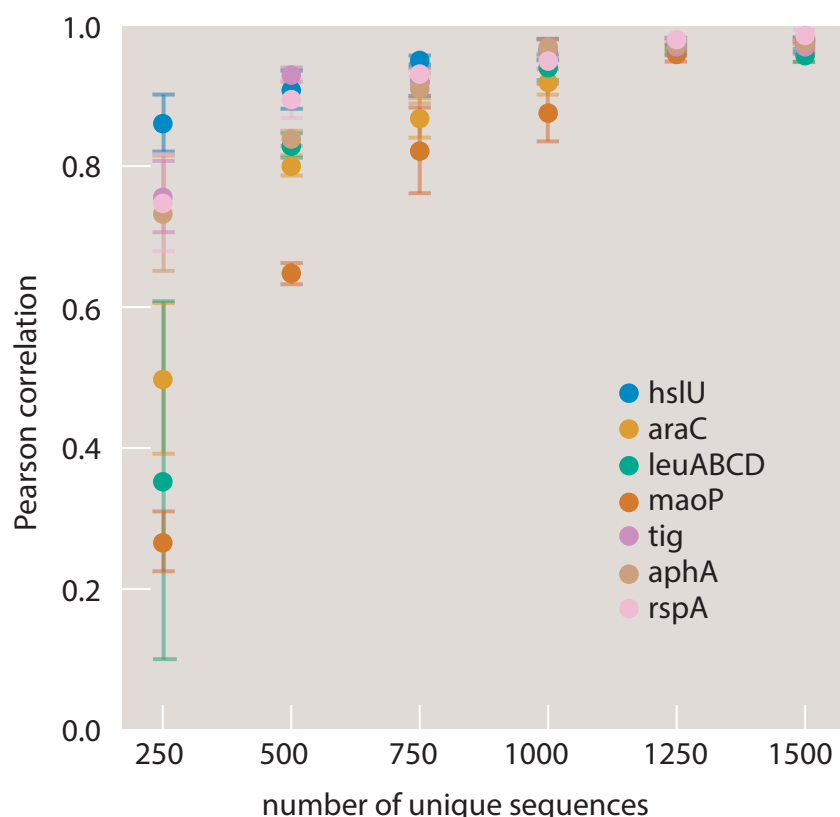


Figure 2.16: Pearson correlation as a function of the number of unique DNA sequences (as explained in Appendix 2.7 Section “Comparison between Reg-Seq by RNA-Seq and fluorescent sorting”). For seven different genes, we studied how the number of mutated DNA sequences affects the reproducibility of our MCMC inference models. As the number of unique sequences increases, so too does the Pearson correlation value, approaching 1.0.

copy number \approx plasmid copy number, there is a corresponding under-estimation of binding energies. Prior work from our laboratory was precisely aimed at rigorously predicting and measuring this effect (Weinert et al., 2014). In that study, we demonstrated how to control this effect, wherein transcription factor copy number \approx plasmid copy number, in a parameter-free manner. However, to mitigate this effect in future studies, we plan to use genome-integrated libraries, rather than plasmid-based expression.

The plasmid used in our experiments is derived from pUA66, which contains a pSC101 origin of replication (Zaslaver et al., 2006). The copy number of plasmids with a pSC101 origin is, in log phase, approximately 3 or 4 (Lutz and Bujard, 1997).

Transcription factor name	Glucose	LB
FNR	609	1101
YieP	158	261
YciT	82	104
NsrR	872	136
LexA	560	1027
DeoR	26	34
CRP	2048	3450
YdfH	96	154
ArcA	3367	5464
Zur	70	130
GlpR	75	145
PhoP	2967	3132
HNS	22541	47133
StpA	6863	5241
DicA	20	25
YgbI	2	6
XylR	1	8

Table 2.7: Global, absolute quantification for most transcription factors identified in this study, as determined for *E. coli* K12 grown in both glucose (5 g/L concentration in M9 minimal media) and LB medias. The values in this table are reprinted from Schmidt et al., 2016 Supplemental Table S6.

We have not independently assessed the copy number of the plasmid used in this study.

The absolute copy number of thousands of proteins in *E. coli* have been determined using whole-proteome LC-MS. Specifically, a 2016 study that provides the absolute quantification for roughly 55 percent of predicted proteins in the *E. coli* K12 proteome (see Supplementary Table S6) (Schmidt et al., 2016). For those transcription factors that were quantified in that study, and also show up in our Reg-Seq experiments, we provide their absolute quantification in *E. coli* K12 for both glucose and LB growth media in Appendix 2.8 Table 2.7.

For most transcription factors, the copy number as determined by LC-MS is much greater than the expected, low copy number of the plasmid used in this study, thus mitigating the concern that the limited availability of a transcription factor could impact gene expression.

There are a few transcription factors that have copy number on the order of the plasmid copy number, however, including XylR, DicA, and YgbI. Prior work from our group (Weinert et al., 2014) has explored how gene expression behaves in the regime where transcription factor copy number is \approx plasmid copy number. Here, we will discuss the case of simple repression to demonstrate how the relationship between transcription factor and plasmid copy number could impact our results. The standard thermodynamic model for gene expression under simple repression with a weak promoter, as described by Bintu et al., 2005, is

$$C \propto p_{bound} = \frac{\frac{P}{N_{NS}} e^{-\beta \Delta \epsilon_P}}{1 + \frac{R}{N_{NS}} e^{-\beta \Delta \epsilon_R}}, \quad (2.24)$$

where C is a measurement for gene expression level, N_{NS} is the number of nonspecific DNA binding sites, P is the number of RNAP, and R is the number of repressors. $\Delta \epsilon_R$ and $\Delta \epsilon_P$ represent the difference in the repressor binding energy and RNAP binding energy between the specific binding site and the averaged nonspecific genomic background respectively. Weinert et al., 2014 demonstrated experimentally that, in the presence of multiple target binding sites, such as from a multi-copy plasmid, the gene expression level can be described by a very similar functional form to equation (2.24), namely,

$$C \propto p_{bound} = \frac{\lambda_P e^{-\beta \Delta \epsilon_P}}{1 + \lambda_R e^{-\beta \Delta \epsilon_R}}, \quad (2.25)$$

where λ_P and λ_R are the fugacity of RNAP and the repressor and describe the relative availability of RNAP or repressor as a function of plasmid copy number, transcription factor copy number, and binding site strength. The presence of additional plasmid copies does weaken the effect of repressor binding when the repressor copy number is \approx plasmid copy number. Thus, our information footprint calculations will be affected and the information signature of binding sites such as YgbI, DicA, or XylR will be decreased.

For transcription factor-binding site interactions that are sufficiently weak, together with a low transcription factor copy number, the effect of having multiple plasmids expressed in a cell could cause us to have a false negative, and thus miss the presence of a binding site. However, the Reg-Seq method does not claim to capture every regulatory feature for a given promoter, as the activity of some transcription factors is induced only in certain growth conditions, we use a finite, 160 bp mutation window

that may miss "regulation at a distance", and the presence of extremely weak and nonspecific binding sites may cause Reg-Seq to "miss" some transcription factors (indeed, for the *bdcR* promoter, the GlcR binding site is outside of the mutagenized region and so is not observed). The effect of additional plasmids within the cellular confines will always decrease the fugacity in equation (2.25), as an increase in the number of sites competing for a limited pool of transcription factors will decrease the relative availability of those transcription factors. As a result, the effect on gene expression of a given transcription factor will always lessen in the presence of additional plasmids. This means that, while multi-copy plasmids can introduce false negatives into Reg-Seq, it will not introduce false positives. Additionally, we see empirically that, even for the lowest copy transcription factor for which we have a measurement, XylR (≈ 1 copy per cell), we can identify its transcription factor binding site. In Appendix 2.7 Figure 2.14, 2 (previously known) XylR sites are identified for the *xylA* promoter, and 2 (previously known) XylR sites are identified in the *xylF* promoter.

Finally, the energy matrices, which are a quantitative output of the Reg-Seq experiment, will be unaffected by the presence of multi-copy plasmids. As discussed in Appendix 2.8 Section Energy matrix inference, energy matrix inference relies on calculating the mutual information between the energy predictions of the model and the experimental data. Mutual information is invariant under transformations to the input variables that do not affect their rank order. While the presence of multiple plasmid copies will affect the fugacity in equation (2.25), and so will also affect any quantitative prediction of gene expression, a weaker repressor binding site will still be predicted to have higher gene expression than a stronger repressor binding site, regardless of the relative availability of the transcription factor. The rank-order is always preserved and so the presence of a multi-copy plasmid will not change the mutual information between model predictions and experimental data. As a result, the final inference of energy matrices will remain the same.

Energy matrix inference

Energy matrices in this experiment are of the form shown in Appendix 2.8 Table 2.8,

where each entry gives the energy contribution from a base pair at a given location. As an example from Appendix 2.8 Table 2.8, an A at position 1 would give a total energy contribution of -0.01 (A.U.). All energy matrices used in our analysis are

pos	A	C	G	T
0	-0.01	-0.01	-0.01	0.03
1	0.002	0.05	-0.06	0.008
2	-0.0002	-0.04	0.008	0.03
3	-0.02	0.02	-0.01	0.01

Table 2.8: Example energy matrix. This matrix is in arbitrary units, and the process to obtain absolute units (in $k_B T$) is described in Appendix 2.8 Section Inference of scaling factors for energy matrices.

linear energy matrices, where the total energy is the sum of contributions from each base pair. As a result, total binding energy is

$$\text{binding energy} = \sum_{i=1}^L \sum_{j=A}^T \theta_{ij} \cdot \delta_{ij}, \quad (2.26)$$

where δ_{ij} is the Kronecker delta, which takes on a value of 1 if the base at position i is equal to j and is 0 otherwise, L is the length of the binding site, and θ_{ij} is the energy contribution of nucleotide j and position i in arbitrary units. To infer the parameters θ_{ij} in equation (2.26) from the experimental data, we perform Bayesian inference using a MCMC method, which requires us to calculate the likelihood of the model given the experimental data. The likelihood function is difficult to determine, but Kinney et al., 2010 found that, given a large amount of data, the likelihood function is related to the mutual information between energy predictions and data by the equation

$$L(D|\theta) \propto 2^{NI(\mu, E)}, \quad (2.27)$$

where N is the total number of independent sequences, D is the data consisting of sequences and measured sequencing counts, I is the mutual information between gene expression label μ and energy predictions E . μ is a discrete variable that characterizes the gene expression level as described in equation (2.3) in the main text. We can calculate mutual information using the formula for mutual information between a continuous and a discrete variable, namely,

$$I(\mu, E) = \int_{-\infty}^{\infty} dE \sum_{\mu=0}^1 p(\mu, E) \log_2 \left(\frac{p(\mu, E)}{p(E)p(\mu)} \right). \quad (2.28)$$

$\mu = 0$	$\mu = 1$	Energy ($k_B T$)
5	23	0.05
5	3	0.008
11	11	0.09
12	3	-0.03
2	0	0.03
8	7	-0.07
7	3	-0.04

Table 2.9: Example dataset with energy predictions. Energy predictions are made by applying the example energy matrix in Appendix 2.8 Table 2.8 to the example dataset in Appendix 2.8 Table 2.6 according to equation (2.26).

While equation (2.28) is used for continuous energy predictions, there are only N total sequences, and so only N discrete energy predictions. For a simple example of calculating the joint probability distribution $p(\mu, E)$, consider the hypothetical dataset of only 4 nucleotides in Appendix 2.8 Table 2.6. We first predict the binding energy of each of the example sequences, shown in Appendix 2.8 Table 2.9.

We use kernel density estimation with kernel width of 4% to estimate the true joint distribution $p(\mu, E_{smooth})$ from the data contained in the joint distribution in the matrix in Appendix 2.8 Table 2.9. This process estimates an underlying continuous distribution from a discrete set of energy predictions. The details of kernel density estimation can be found in Hastie, Tibshirani, and Friedman, 2009. We can do the final calculation of the mutual information by splitting the smoothed joint distribution into 500 energy "bins" z and calculating

$$I(\mu, E) = \sum_{z=1}^{500} \sum_{\mu=0}^1 p(\mu, E_z) \log_2 \left(\frac{p(\mu, E_z)}{p(E_z)p(\mu)} \right). \quad (2.29)$$

With the ability to calculate the likelihood of an energy matrix model, MCMC can be used to infer the posterior distribution for our model. First a random matrix model is generated. The model is perturbed and the new model is accepted or rejected based on the Metropolis-Hastings algorithm (Patil, Huard, and Fannesbeck, 2010). After an initial burn in period of 60000 steps, iterations are saved every 60 steps. A total of 600000 iterations are performed. This procedure is performed twice for each model, and if inferred models do not have a Pearson correlation coefficient of

0.99 or higher they are discarded and computed again. A complete overview of the computational pipeline can be found at the GitHub wiki page.

Inference of scaling factors for energy matrices

For the majority of energy matrices reported in our work, the results are given in arbitrary units. This is a direct result of using the method of Kinney et al., 2010 to infer our matrices. The method appeals to information theory to write an "error-model-averaged" likelihood function for a given model. The likelihood function is given in equation (2.27). A property of mutual information is that it is invariant to changes in the input variables as long as those transformations do not affect the rank-order of those variables. As a result, we can scale the energy predictions by any constant without changing the likelihood of the model, which means that in the case of simple linear models for transcription factor binding we cannot assign absolute units to energy matrix values. When we widen our view to considering promoter regions rather than single binding sites we can overcome this drawback. Using thermodynamic modeling as outlined in Bintu et al., 2005 we can predict the gene expression from any given transcriptional architecture. In the case a thermodynamic model of simple repression the expression is given by

$$C \propto p_{bound} = \frac{\frac{P}{N_{NS}} e^{-\beta \Delta \epsilon_P}}{1 + \frac{P}{N_{NS}} e^{-\beta \Delta \epsilon_P} + \frac{R}{N_{NS}} e^{-\beta \Delta \epsilon_R}}, \quad (2.30)$$

where C is a measurement for expression, P is the number of RNAP, R is the number of repressors and N_{NS} is the number of nonspecific binding sites. $\Delta \epsilon_R$ and $\Delta \epsilon_P$ represent the difference in the repressor binding energy and RNAP binding energy between the specific binding site and the averaged nonspecific genomic background respectively. As we use linear energy matrix models as described in Appendix 2.8 Section Energy matrix inference, $\Delta \epsilon_R$ and $\Delta \epsilon_P$ will be given by equation (2.26). In these cases the overall rank order of gene expression predictions will change if you scale the energy matrix, and so the absolute units can be determined (Kinney and Atwal, 2014). Equation (2.30) is a more complicated and non-linear functional form for predicting C than a simple linear binding model, and has a correspondingly more difficult to sample posterior. To address complications in the inference, we first only use the non-linear fits to fix overall scale and wild type energy for energy matrices rather than fit all parameters in this way. In other words we use the standard fitting procedure to find the θ_{ij} in the equation (2.26) using the standard MCMC procedure.

gene	growth	scaling factor A
<i>tff-rpsB-tsf</i>	Heat shock	$-8.1 k_B T$
<i>tig</i>	Heat shock	$-26.3 k_B T$
<i>yjjJ</i>	Heat shock	$-11.3 k_B T$
<i>bdcR</i>	Heat shock	$-9.9 k_B T$
<i>fdhE</i>	Anaerobic growth	$-6.34 k_B T$
<i>ykgE</i>	Arabinose	$-12.1 k_B T$
<i>dicC</i>	Arabinose	$-15.1 k_B T$
<i>rspA</i>	Arabinose	$-5.5 k_B T$

Table 2.10: Scaling factors to convert arbitrary units to absolute units in $k_B T$. Growth conditions indicate the energy matrix and dataset used in the fit. In some growth conditions additional regulatory features will be present, meaning the specific condition used for inference is important.

The binding energy matrices can be written $A \cdot \theta_{ij} + B$ where A is a constant that scales the matrix from arbitrary units to absolute units ($k_B T$) and B is an additive constant that relates to the wild type energy. We fit the constants A and B for the transcription factor binding energy using the thermodynamic model in equation (2.30).

While we can in principle fit thermodynamic models to any given architecture, these models are non-linear and, due to numerical difficulties, unreliable for sufficiently complex models. We only use this method on examples of simple repression or activation without more than one prominent RNAP model, whose transcription factor binding site does not overlap significantly with RNAP -10 or -35 sites. The scaling factors we discovered are given in Appendix 2.8 Table 2.10.

We perform the inference using parallel tempering MCMC, where multiple chains are run in parallel with different "temperatures". High temperature chains widely explore parameter space, escaping any local optima, while low temperature chains optimize locally. The current parameter values of the chains are exchanged periodically. The fitting procedure is done using the emcee ensemble sampler (Goodman and Weare, 2010) with 10 temperatures ranging from 1 to 10000 on a log scale.

Examination of promoters for which no RNAP site was found

We failed to find an RNAP site for 18 promoter sequences. In order to understand these sequences in more detail we examine the sequences within 50 bases of the TSS for the 18 genes in question for sequences which resemble the known consensus RNAP binding site. For this comparison we use the σ^{70} consensus binding sequence

$-^{35}\text{TTGACA}$ - spacer sequence - TGNTATAAT^{-7} (where the superscripts $-^{35}$ and $-^7$ indicate the position relative to the TSS). The consensus sequence we use for comparison contains the extended -10 element, consisting of a TG at bases -15 and -14 as we have found those to be important for gene expression in our study. The spacer length is between 15 and 13 bases (the typically reported spacer length is between 18 and 16 but this does not include the extended minus 10 element). The consensus sequence for the heat shock σ factor was used for the promoter *yajL*.

Previously, to analyze RNAP sites, we have examined energy matrices produced by Reg-Seq. Now we add an examination of wild type sequences. For each promoter, we found the best match to the consensus site, namely the sequence with the fewest mutations compared to the consensus sequence. We use the number of mutations as a measure of how well the site resemble consensus. We find that 16 out of 18 promoters have at least 5 mutations in the sequence that most closely resembles RNAP, one promoter has 4 mutations, and the last has three mutations. To put these numbers into context, Brewster, Jones, and Phillips, 2012 measured the RNAP binding energies of several RNAP binding site mutants. Mutations away from the strongest sequence tested (*lacUV5*, which is 2 mutations away from consensus) yields a change in binding energy of $\approx 1 - 2 k_B T$. If the promoters are constitutive, then (in the weak promoter approximation) expression level will be proportional to $e^{-\beta \Delta \epsilon_P}$ where $\Delta \epsilon_P$ is the RNAP binding energy relative to the nonspecific background. Therefore, as an approximation, a sequence with 3 mutations would be predicted to be 3 – 10 fold weaker than a "strong" RNAP site, and as such could be said to show a resemblance to the consensus RNAP site. However, 16 out of 18 of these promoter regions have, at best, extremely weak RNAP sites. It is important to note however, that even extremely weak RNAP sites often transcribe, especially when aided by activators. We do not intend to claim that RNAP does not bind to these promoter regions, merely that we do not detect it in our experiment. In fact, while the RNAP sites are weak, there is experimental evidence in EcoCyc of some level of transcription for 9 out of 18 promoters.

TOMTOM motif comparison

In some cases, we used an alternative approach to mass spectrometry to discover the transcription factor identity regulating a given promoter based on sequence analysis using a motif comparison tool. TOMTOM (Gupta et al., 2007) is a tool that uses a statistical method to infer if a putative motif resembles any previously discovered motif in a database. It accounts for all possible offsets between the motifs.

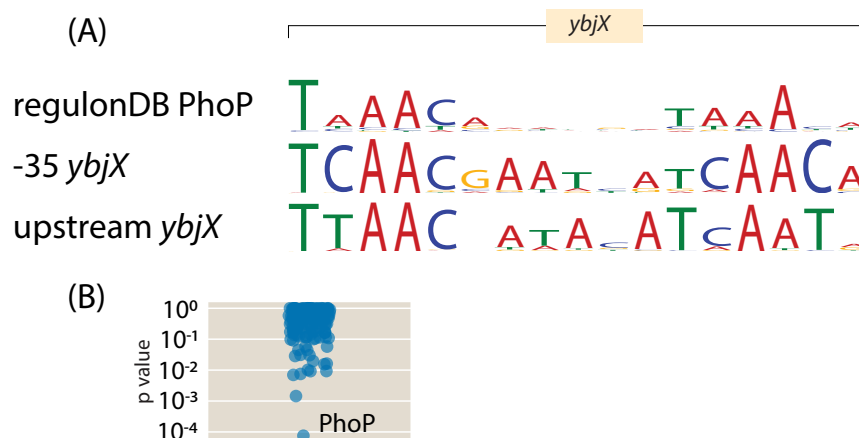


Figure 2.17: Motif comparison using TOMTOM for the two PhoP binding sites in the *ybjX* promoter. Searching our energy motifs against the RegulonDB database using TOMTOM allowed us to guide our transcription factor knockout experiments. Here we show the sequence logos of the PhoP transcription factor from RegulonDB (top) and the ones generated from the *ybjX* promoter energy matrix. E-value = 0.01 using Euclidean distance as a similarity matrix.

Moreover, it uses a suite of metrics to compare between motifs such as Kullback-Leibler divergence, Pearson correlation, Euclidean distance, among others. All TOMTOM analyses in Reg-Seq utilize Euclidean distance. The method calculates a p-value under the null hypothesis that the two compared motifs are independently drawn from the same underlying distribution probability distribution.

We performed comparisons of the motifs generated from our energy matrices to those generated from all known transcription factor binding sites in RegulonDB. Appendix 2.8 Figure 2.17 shows a result of TOMTOM, where we compared the motif derived from the -35 region of the *ybjX* promoter and found a good match with the motif of PhoP from RegulonDB.

The information derived from this approach was then used to guide some of the transcription factor knockout experiments, in order to validate its interaction with a target promoter characterized by the loss of the information footprint. Furthermore, we also used TOMTOM to search for similarities between our own database of motifs, in order to generate regulatory hypotheses in tandem. This was particularly useful when looking at the group of GlpR binding sites found in this experiment.

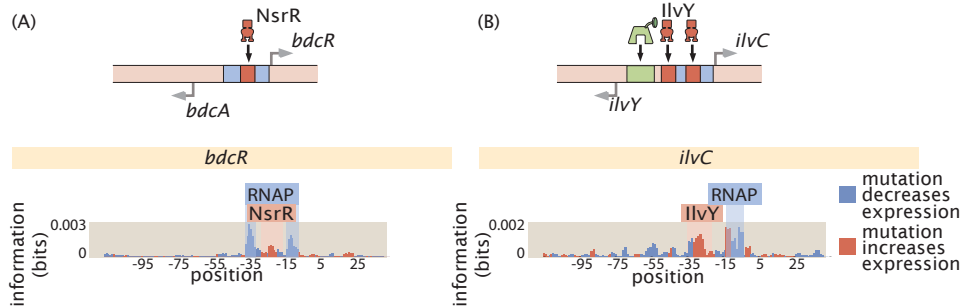


Figure 2.18: Two cases in which we see transcription factor binding sites that we have found to regulate both of the two divergently transcribed genes. (A) An information footprint and regulatory cartoon for the divergently transcribed *bdcA* and *bdcR* promoters. A single NsrR site regulates both promoters. (B) An information footprint and regulatory cartoon for the *ilvC* and *ilvY* promoters. Both promoters are repressed by IlvY when grown without acetolactate. Only the IlvY site is labeled on the information footprint.

2.9 Supplementary information: Additional Results

Binding sites regulating divergent operons

In addition to discovering new binding sites, we have discovered additional functions of known binding sites. In particular, in the case of *bdcR*, the repressor for the *bdcA* gene, which is transcribed from the same promoter in the opposite direction of transcription (Partridge et al., 2009), is also shown to repress *bdcR* in Appendix 2.9 Figure 2.18(A). Similarly in Appendix 2.9 Figure 2.18(B) IlvY is shown to repress *ilvC* in the absence of inducer. Divergently (transcription in opposite directions from the same promoter) transcribed operons that share regulatory regions are plentiful in *E. coli*, and although there are already many known examples of transcription factor binding sites regulating several different operons, there are almost certainly many examples of this type of transcription that have yet to be discovered.

In the case of *ilvC*, IlvY is known to activate *ilvC* in the presence of inducer. However, we now see that it also represses the promoter in the absence of that inducer. The production of *ilvC* is known to increase by approximately a factor of 100 in the presence of inducer (Rhee, Senear, and Hatfield, 1998). The magnitude of the change is attributed to the cooperative binding of two IlvY binding sites, but the lowered expression of the promoter due to IlvY repression in the absence of inducer is also a factor.

Comparison of results to RegulonDB

One area in which our work can be compared to current repositories of regulatory information such as RegulonDB is in comparing the prevalence of different regulatory architectures in the database to Reg-Seq. Appendix 2.9 Figure 2.19 shows the prevalence of each type of architecture (not including architectures more complex than 2 activators and 2 repressors), and shows how simpler architectures are more common in both cases.

Another point of comparison between RegulonDB and Reg-Seq can be found in comparing sequence motifs from Reg-Seq to those generated from binding sites in RegulonDB. This can often produce useful results, such as in Appendix 2.8 Section “TOMTOM motif comparison”. For other cases the data used to generate the RegulonDB motifs can be lacking. We believe the GlpR motif in RegulonDB highlights some of the issues with using the reported motifs in RegulonDB to predict binding preference. First, there are only 4 promoters regulated by GlpR, with a total of 17 binding sites for GlpR in RegulonDB. 9 of these binding sites differ by 9 mutations or more from the consensus site (out of 22 total base pairs). 9 mutations is more than even the weak O3 operator for LacI. We do believe that a relatively low number of weakly conserved binding sites likely do not reveal quality sequence logos for a binding site, especially as compared to Reg-Seq which constructs sequence logos from over a thousand promoter variants. Generation of such sequence motifs is a point on which we believe Reg-Seq can improve the current status of regulatory knowledge.

Explanation of included binding sites

This section is intended to clarify cases in which the regulatory cartoon or the displayed "expected" binding sites differs from what can be found in RegulonDB or EcoCyc. The primary reason for these discrepancies is that our experiment only targets a 160 base pair mutation window. Some known binding sites will be outside of this window. Additionally, while some genes are known to be regulated by a specific transcription factor, the exact location of that transcription factor's binding site is unknown and so we cannot be certain during the design of the 160 base pair mutagenized window whether or not the transcription factor binding site will be present in our experiment. The locations of the TSS selected in this experiment can be found in Supplementary File 1. Additionally, some transcription factors are known to only be active under certain growth conditions. Information footprints are

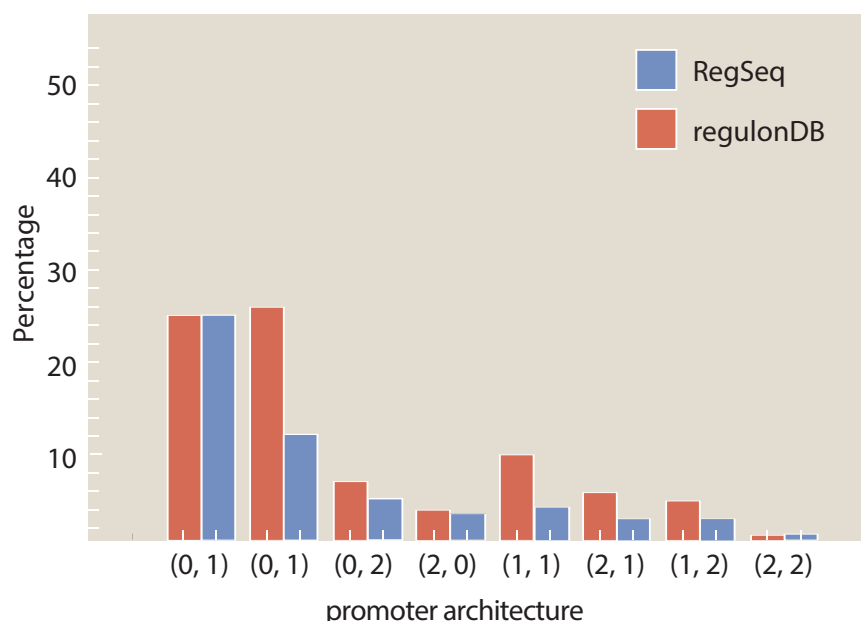


Figure 2.19: A comparison of the types of architectures found in RegulonDB (Santos-Zavaleta et al., 2019) to the architectures with newly discovered binding sites found in the Reg-Seq study. For each type of architecture, labeled as (number of activators, number of repressors), the fraction that architecture comprises of the total number of operons is given both for the data found in RegulonDB and from the results of the Reg-Seq experiment.

depictions of the regulatory information for a specific growth condition; accordingly, not all transcription factor binding sites can be identified using a single growth condition. Throughout the main text and SI, however, we depict regulatory cartoons with their full milieu of transcription factors (based on experiments performed in multiple growth conditions).

When devising this study, we sought to test the reliability of the Reg-Seq method by testing experimentally-validated transcription factor binding sites, as reported by EcoCyc or RegulonDB, to assess our ability to recapitulate prior experiments. EcoCyc labels some transcription factor binding sites as "low-evidence" in their database, most of which were identified via sequence motif matching. We have repeatedly observed that transcription factor binding sites identified from sequence matching are unreliable in relation to the empirical data collected in our experiment, and so we choose not to include them in the set of "gold standard" genes which were used for this purpose of assessing Reg-Seq's accuracy.

All of our "gold standards" are genes for which there is high quality experimental

evidence of their transcriptional regulation and the location of related transcription factor binding sites and, again, they were used to evaluate the false negative rates of our experiment. In those cases where the binding sites are either 'low-evidence' according to EcoCyc, the location of a binding site is not known, a gene is only actively transcribed in certain or unknown growth conditions, or the binding site location is outside of the 160 bp mutagenized region, we do not include them in the list of sites we use to test our method even though they appear as binding sites in RegulonDB or EcoCyc. Regulatory features that are not transcription factors, including regulatory RNAs, are also not labeled in our reported results.

Accordingly, in some cases, the regulatory cartoons or architectures we present in this study may appear to be incomplete relative to previous reports of promoter architectures. For each gene below, we explain these discrepancies. This section is intended to explain why annotations on information footprints or regulatory cartoons do not match what is seen in RegulonDB or EcoCyc.

sdiA

sdiA is known to be regulated by both Nac as well as CsrA (which has two binding sites), the CsrA sites are downstream of the mutated region and the location of the Nac binding site is unknown. Thus, none of these binding sites are reported in our regulatory architectures for this gene.

yqhC

yqhC is known to be regulated by GlaR, but the location of this binding site is unknown. As a result, we were unable to identify this binding site in our analysis, and the architecture for *yqhC* is listed in this study as (0,0).

bdcR

bdcR is known to be regulated by GlaR, but this binding site is outside of the targeted mutation window of 160 bp. A known binding site for NsrR is included within the 160 bp region, but it was not previously known to regulate *bdcR*; the binding site for NsrR is included as a new discovery as shown in Section Binding sites regulating divergent operons.

aegA

aegA has a predicted CRP binding site, but the location of this binding site is unknown and it is also listed as low-evidence in EcoCyc. As a result, the site is not included within this study's analysis.

hicB

The CRP site associated with *hicB* is cited as low-evidence in EcoCyc and the HicB binding site is outside of the 160 base pair mutated region. As a result, neither site is included in this study.

rplKAJL-rpoBC

The known RplA binding site for this operon is outside of the targeted, 160 base pair mutation window. As a result, the RplA site is not included in this study.

tff-rpsB-tsrf

RpsB is not contained in the mutated region. Additionally, the nearby predicted Mar-Sox-Rob binding site is listed as low-evidence in EcoCyc and is also not directly predicted to regulate *tff-rpsB-tsrf*, even though it may be present within the region. As a result, neither site is included in this study.

yodB

GlaR is known to regulate *yodB*. However, the location of this binding site is unknown. As a result, we do not include the GlaR binding site in our reported regulatory architecture for this gene.

maoP

HdfR is known to regulate *maoP*. However, the location of the binding site is unknown. Additionally, the HdfR site is listed as low-evidence in EcoCyc. During the Reg-Seq experiment, however, we confirmed the presence of the low-evidence HdfR site with a gene knockout and located the binding site position. Thus, we include it in all regulatory cartoons and report the HdfR site in our discoveries.

poxB

MarA and Sox have low-evidence binding sites in the mutagenized region. There is also a low-evidence site for Cra with an unknown binding location. As a result, neither site is included in the reported regulatory architectures in this study.

mscM

While there is a known CpxR binding site for *mscM*, the binding site exists outside of the mutagenized region. As a result, it is not included in the reported regulatory architectures in this study.

tar

There is a low-evidence FNR site for *tar*. Its location is unknown. For both of these reasons we do not include the binding site in our reported regulatory architectures for this gene.

dpiBA

While there are 10 total binding sites for *dpiBA*, including an FNR site. However, the only ones that are known to regulate the particular TSS we chose (at position 652172 in *E. coli*) are 2 DcuR sites and a (low-evidence) NarL site. DcuR is induced by growth conditions like succinate or fumarate, neither of which were tested in this study. As a result, none of the sites are included in this study.

araAB

There are a total of 5 AraC binding sites and one CRP binding site that regulate *araAB*. However, the three furthest upstream AraC binding sites are outside of the 160 bp mutagenized region, and so only 2 AraC sites and one CRP site is included in the reported regulatory architecture in this study.

xylF

There are two XylR sites, as well as 3 low-evidence Fis sites that regulate *xylF* in the mutagenized region. There is also a low-evidence CRP site outside the mutagenized region. Only the 2 XylR sites are included in the reported regulatory architectures, as the remaining sites are low-evidence or outside the mutagenized region.

xylA

There are 2 XylR sites, 2 AraC, and a CRP site that regulates *xylA*. In our analysis, we utilize a growth condition containing xylose and arabinose. Under growth with xylose, XylR will bind DNA and activate expression. Under growth with arabinose, AraC will not bind DNA. We would only expect to see two XylR sites and a CRP site under growth in xylose and arabinose, so we only include these sites in our study.

dicB

DicA has a low-evidence repressor binding site for *dicB*. Additionally the binding location is unknown, and so we do not include the binding site in the reported regulatory architecture.

xapAB

XapR has two low-evidence binding sites. The binding site furthest upstream is outside of the 160 bp mutagenized region. As the remaining site is low-evidence, it is not included in our reported regulatory architectures.

ilvC

There are two IlvY binding sites for *ilvC*. IlvY is known to be induced by acetolactate and activated in its presence. We do not utilize this growth condition in this experiment, however, nor do we include the two IlvY binding sites in our "gold standard" experimental analysis. We find that IlvY acts as a repressor when grown in other growth conditions. As repressor activity at these sites was not previously reported, we include this in our list of new discoveries.

asnA

There are 4 low-evidence AsnC binding sites in the mutated region. As they are low-evidence, however, we do not include these binding sites in the reported regulatory architectures for this gene.

idnK

While there are 3 GntR sites, a CRP site, and one IdnR site, they are all low-evidence. As a result, we do not include any of these sites in our reported regulatory architectures.

dinJ

While *dinJ* is regulated by DinJ-YafQ and LexA, they are both outside of the mutagenized window. As a result, neither are included in our reported regulatory architectures.

yjiY

yjiY is regulated by both BtsR and CRP. However, CRP is outside of the mutagenized window and so CRP is not included in our reported regulatory architectures.

cra

cra is regulated by a low-evidence binding site of PhoB. The location of the binding site is not known, however. As a result, the site is not included in the reported regulatory architecture.

uvrD

uvrD is regulated by a low-evidence binding site for LexA. This binding site is not included in the reported regulatory architectures for this study.

znuCB

There are binding sites for Zur and OxyR in the mutagenized region for *znuCB*. OxyR is known to act as an activator under oxidative stress. As we do not utilize an oxidative stress growth condition in this study, we do not include this binding site in the reported regulatory architectures for this study.

znuA

There are binding sites for Zur and OxyR in the mutagenized region for *znuA*. The OxyR binding site is outside of the mutagenized region. Only the Zur binding site is included in our reported regulatory architectures.

pitA

There is a low-evidence binding site for FNR in the mutagenized region. The location of this binding site, however, is unknown. Thus, this binding site is not included in our reported regulatory architectures.

ecnB

There is a low-evidence OmpR binding site for *ecnB*. The binding site is not included in our reported regulatory architectures.

lacZYA

The mutagenized region extends from the TSS (the primary TSS p1) to 75 base pairs upstream of the TSS. The location of the mutagenized region excludes the LacI sites, while including a single CRP binding site, a MarA binding site, and two HNS binding sites. The expression from *marA* is expected to be low, as we do not grow the cells in the presence salicylate or antibiotic stress and so we do not expect to observe the MarA site. In fact, the precursor of the Reg-Seq experiment, Sort-Seq, mutagenized and studied the same 75 base pair region, and only observed binding by CRP (Kinney et al., 2010). As such, we only include CRP in Table 2, the regulatory cartoons, or the analysis of false positives and false negatives.

leuABCD

There is a binding site for LeuO regulating *leuABCD*. The site is low-evidence and also has no known binding location. As a result, the site is not included in our reported regulatory architectures.

arcA

There is a binding site for FNR within the mutagenized region listed as "low-evidence" in EcoCyc. We find substantial additional evidence for the presence of the FNR binding site. As such, we include the site in Table 2.2 as an "Identified Binding Site."

relBE

The *relBE* promoter contains 4 RelBE binding sites and 2 RelB binding sites in EcoCyc and RegulonDB. While the all 4 RelBE sites are listed as high evidence, Belliveau et al., 2018 mutagenized the RelBE promoter and did not identify binding in the furthest downstream or furthest upstream binding sites. Also, the original identification of the RelBE binding sites presented (Li et al., 2008), claims that the furthest upstream and downstream sites are only identified by similarity to consensus

sequence. As a result only 2 of the RelBE and 2 of the RelB sites are included in this study.

marR

The *marR* promoter contains a CpxR, CRP, Cra, and AcrR in EcoCyc that are not included in the "gold standard" analysis or Table 2. Belliveau et al., 2018 performed mutagenesis experiments on the *marR* promoter and did not identify these additional sites and so they have been excluded.

2.10 Supplementary information: Key Resource Table

(Table included on the following pages.)

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Cell line (<i>Escherichia coli</i>)	<i>E. Coli</i> K12	<i>E. Coli</i> Stock Center		
Cell line (<i>Escherichia coli</i>)	<i>E. Coli</i> ΔYieP	<i>E. Coli</i> Stock Center		
Cell line (<i>Escherichia coli</i>)	<i>E. Coli</i> ΔGlpR	<i>E. Coli</i> Stock Center		
Cell line (<i>Escherichia coli</i>)	<i>E. Coli</i> ΔArcA	<i>E. Coli</i> Stock Center		
Cell line (<i>Escherichia coli</i>)	<i>E. Coli</i> ΔLrhA	<i>E. Coli</i> Stock Center		
Cell line (<i>Escherichia coli</i>)	<i>E. Coli</i> ΔPhoP	<i>E. Coli</i> Stock Center		
Cell line (<i>Escherichia coli</i>)	<i>E. Coli</i> ΔHdfR	<i>E. Coli</i> Stock Center		
Strain (<i>Escherichia coli</i>)	<i>E. Coli</i> ΔGlpR in K12 strain	This paper		Knockout transferred to <i>E. coli</i> K12
Strain (<i>Escherichia coli</i>)	<i>E. Coli</i> ΔArcA in K12 strain	This paper		Knockout transferred to <i>E. coli</i> K12
Strain (<i>Escherichia coli</i>)	<i>E. Coli</i> ΔLrhA in K12 strain	This paper		Knockout transferred to <i>E. coli</i> K12
Strain (<i>Escherichia coli</i>)	<i>E. Coli</i> ΔPhoP in K12 strain	This paper		Knockout transferred to <i>E. coli</i> K12
Strain (<i>Escherichia coli</i>)	<i>E. Coli</i> ΔHdfR in K12 strain	This paper		Knockout transferred to <i>E. coli</i> K12
commercial assay or kit	RNEasy Mini kit	Qiagen	Cat.: 74104	
chemical compound, drug	Q5 Polymerase	Qiagen	Cat.: M0491L	
chemical compound, drug	qPCR master mix	QuantaBio	Cat.: 101414-166	
chemical compound, drug	Lysyl Endopeptidase	Wako Chemicals	Cat.: 125-05061	
chemical compound, drug	RNAprotect bacteria reagent	Qiagen	Cat.: 76506	
other	streptavidin coated dynabeads	Thermo Fisher	Cat.: 65601	
software, algorithm	mpathic	Kinney Lab	Ireland et al. 2016	
software, algorithm	FastX	Hannon Lab	RRID:SCR_005534	
software, algorithm	FLASH	CBCB	RRID:SCR_005531	

continued on following page

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
sequence based reagent	oligo Pool	Twist Bioscience		
sequence based reagent	fwd oligo 101	IDT		TTCGTCTTCACCTCGAGCAC GCTTATTCGTGCCGTGTTAT
sequence based reagent	fwd oligo 102	IDT		TTCGTCTTCACCTCGAGCAC TTTGCTTCAGTCAGATTCGC
sequence based reagent	fwd oligo 103	IDT		TTCGTCTTCACCTCGAGCAC GTCGAGTCCTATGTAACCGT
sequence based reagent	fwd oligo 104	IDT		TTCGTCTTCACCTCGAGCAC GTAAGATGGAAGCCGGGATA
sequence based reagent	fwd oligo 105	IDT		TTCGTCTTCACCTCGAGCAC GGTGTCGCAACATGATCTAC
sequence based reagent	fwd oligo 106	IDT		TTCGTCTTCACCTCGAGCAC GTGCTAAGTCACACTGTTGG
sequence based reagent	fwd oligo 107	IDT		TTCGTCTTCACCTCGAGCAC TCTAAACAGTTAGGCCAGG
sequence based reagent	fwd oligo 108	IDT		TTCGTCTTCACCTCGAGCAC GTCTTTATACTTGCCTGCCG
sequence based reagent	fwd oligo 109	IDT		TTCGTCTTCACCTCGAGCAC CACCGCGATCAATACAACCTT
sequence based reagent	fwd oligo 110	IDT		TTCGTCTTCACCTCGAGCAC TTCGGATAGACTCAGGAAGC
sequence based reagent	fwd oligo 111	IDT		TTCGTCTTCACCTCGAGCAC CCATTGATAGATTCGCTCGC
sequence based reagent	fwd oligo 112	IDT		TTCGTCTTCACCTCGAGCAC TTTTCTACTTTCCGGCTTGC
sequence based reagent	fwd oligo 113	IDT		TTCGTCTTCACCTCGAGCAC ATGACTATTGGGGTCGTACC
sequence based reagent	fwd oligo 114	IDT		TTCGTCTTCACCTCGAGCAC TCGACAATAGTTGAGCCCTT

continued on following page

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
sequence based reagent	fwd oligo 115	IDT		TTCGTCTTCACCTCGAGCAC GAGCCATGTGAAATGTGTGT
sequence based reagent	fwd oligo 116	IDT		TTCGTCTTCACCTCGAGCAC CGTATACGTAAGGGTTCCGA
sequence based reagent	fwd oligo 117	IDT		TTCGTCTTCACCTCGAGCAC TTATGATGTCCGGATACCCG
sequence based reagent	fwd oligo 118	IDT		TTCGTCTTCACCTCGAGCAC TCTTAGAAATCCACGGGTCC
sequence based reagent	rev oligo 101	IDT		TGTAACGACGGCCAGTGACT AGCGCTGAGGAGAAGCCTAATA GGGCACAGCAATCAAAAGTA
sequence based reagent	rev oligo 102	IDT		TGTAACGACGGCCAGTGAGG AGCGCTGAGGAGAAGCCTAATA CCGGGATTCAGTGATTGAAC
sequence based reagent	rev oligo 103	IDT		TGTAACGACGGCCAGTGAGT CCCGCTGAGGAGAAGCCTAATA TGAAGATATGACGACCCCTG
sequence based reagent	rev oligo 104	IDT		TGTAACGACGGCCAGTGACC GACGCTGAGGAGAAGCCTAATA TTCCACAGCTCTATGAGGTG
sequence based reagent	rev oligo 105	IDT		TGTAACGACGGCCAGTGATT GGCGCTGAGGAGAAGCCTAATA GCAAACATGACTAGGAACCG
sequence based reagent	rev oligo 106	IDT		TGTAACGACGGCCAGTGAGA TACGCTGAGGAGAAGCCTAATA CCGGGACGAGATTAGTACAA
sequence based reagent	rev oligo 107	IDT		TGTAACGACGGCCAGTGAAC TCCGCTGAGGAGAAGCCTAATA CACGCCAGTTGTGAACATAA

continued on following page

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
sequence based reagent	rev oligo 108	IDT		TGTAAAACGACGGCCAGTGATA CTCGCTGAGGAGAAGCCTAATA CAAAGGCCAAATCAGTTCCA
sequence based reagent	rev oligo 109	IDT		TGTAAAACGACGGCCAGTGACC AACGCTGAGGAGAAGCCTAATA GGTGCATGGGAGGA ACTATA
sequence based reagent	rev oligo 110	IDT		TGTAAAACGACGGCCAGTGAAG GCCGCTGAGGAGAAGCCTAATA TGCATGGGTCTGTCTATTGT
sequence based reagent	rev oligo 111	IDT		TGTAAAACGACGGCCAGTGAAA TTCGCTGAGGAGAAGCCTAATA CTCCTATGCTAGCTCGACTC
sequence based reagent	rev oligo 112	IDT		TGTAAAACGACGGCCAGTGATT GTCGCTGAGGAGAAGCCTAATA ATGGTAAGAAGCTCCCACAA
sequence based reagent	rev oligo 113	IDT		TGTAAAACGACGGCCAGTGATT TACGCTGAGGAGAAGCCTAATA CTATGGTCATTCCCGTACGA
sequence based reagent	rev oligo 114	IDT		TGTAAAACGACGGCCAGTGAAC CGCGCTGAGGAGAAGCCTAATA TAATCGGCTACGTTGTGTCT
sequence based reagent	rev oligo 115	IDT		TGTAAAACGACGGCCAGTGATG GCCGCTGAGGAGAAGCCTAATA TGA CTGATCCTTTAGTCCG
sequence based reagent	rev oligo 116	IDT		TGTAAAACGACGGCCAGTGAGG CCCGCTGAGGAGAAGCCTAATA ACGCTTTGTGTTATCCGATG
sequence based reagent	rev oligo 117	IDT		TGTAAAACGACGGCCAGTGAGG TGCGCTGAGGAGAAGCCTAATA ACCACGGTGGAGTATACATC

continued on following page

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
sequence based reagent	rev oligo 118	IDT		TGTAACACGACGGCCAGTGACA ATCGCTGAGGAGAAGCCTAATA GGCACCAGGTACATATCTCA
sequence based reagent	mRNA rev	IDT		GCAGGGGATAATATTGCCCA
sequence based reagent	fwd sequencing 94	IDT		AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGAT CTGACCTATTAGGCTTCTCCTCAGCG
sequence based reagent	fwd sequencing 95	IDT		AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCT CAGTTATTAGGCTTCTCCTCAGCG
sequence based reagent	fwd sequencing 96	IDT		AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCT TCTATATTAGGCTTCTCCTCAGCG
sequence based reagent	fwd sequencing 97	IDT		AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCT AGAGTATTAGGCTTCTCCTCAGCG
sequence based reagent	fwd sequencing 98	IDT		AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCT GCATTATTAGGCTTCTCCTCAGCG
sequence based reagent	fwd sequencing 99	IDT		AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCT CTTATATTAGGCTTCTCCTCAGCG
sequence based reagent	fwd sequencing 100	IDT		AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCT TAGCTATTAGGCTTCTCCTCAGCG
sequence based reagent	fwd sequencing 101	IDT		AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCT CAAGTATTAGGCTTCTCCTCAGCG

continued on following page

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
sequence based reagent	fwd sequencing 102	IDT		AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCT GTACTATTAGGCTTCTCCTCAGCG
sequence based reagent	fwd sequencing 103	IDT		AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCT TGAATATTAGGCTTCTCCTCAGCG
sequence based reagent	fwd sequencing 104	IDT		AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCT TCGTTATTAGGCTTCTCCTCAGCG
sequence based reagent	fwd sequencing 105	IDT		AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCT ATGCTATTAGGCTTCTCCTCAGCG
sequence based reagent	fwd sequencing 106	IDT		AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCT GTCATATTAGGCTTCTCCTCAGCG
sequence based reagent	fwd sequencing 107	IDT		AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCT CTCATATTAGGCTTCTCCTCAGCG
sequence based reagent	fwd sequencing 108	IDT		AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCT AGTATATTAGGCTTCTCCTCAGCG
sequence based reagent	rev sequencing	IDT		AAGCAGAAGACGGCATACGAGATCGGTCT CGGCATTCCTGCTGAACCGCTCTTCCGAT CTCAAAGCAGGGGATAATATTGCCCA
database database	RegulonDB EcoCyc		RRID:SCR_003499 RRID:SCR_002433	

Table 2.11: Key Resource Table. This table contains the resources needed to replicate the Reg-Seq experiment.

Chapter 3

QUANTITATIVE DISSECTION OF A SINGLE PROMOTER USING RNA-SEQ

The following work is currently being conducted in collaboration with Manuel Razo-Mejia, Tom Röschinger, and Scott Saunders. Nicholas McCarty also conducted key preliminary experiments that served as a foundation for the work included here. What follows is a work in progress for what will ideally be published in the coming months. The primary focus of this chapter is to introduce the motivation for the project and to outline some of the key materials and methods that have been used.

3.1 Motivation

In Chapter 1, we laid out what it means to quantitatively dissect a regulatory region, with the *lac* operon as our gold standard. However, we also laid out the enormous issue of our regulatory ignorance, which was the entire focus of Chapter 2. Now we would like to come full circle: once we have completed the ‘fact finding mission’ of conducting a Reg-Seq experiment in full, can we actually give these newly found regulatory architectures the same predictive treatment that the *lac* operon has now been put through time and time again (as illustrated in Figure 3.1)? To be truly scalable, we would like this quantitative dissection to be sequencing-based, just like the process of regulatory discovery enabled via Reg-Seq.

In the work that is currently in progress, we have a number of goals, in increasing scope and complexity: (1) We would like to be able to first and foremost confirm that we can use RNA-Seq to quantitatively validate what is already known for the *lac* operon. For this, we will use the most minimal ‘library’ of sequences of just the native *lac* repressor operators: O1, O2, and O3. (2) Increasing the complexity to more than just a few native operators, we would like to be able to dissect an entire library of operator mutants, expanding on the work done by Barnes et al., 2019, where each mutant had to be painstakingly cloned independently. For this, we have a library containing over 3000 operator variants, all of which we will be able to test in a single sequencing experiment. (3) And finally, while repeating and building off of the work already done on *lac* serves as a great proving ground, we ultimately want to bring these tools to bare on novel discovered regulation, such as *purR* found by Belliveau et al., 2018 or any of the number of recently discovered architectures from

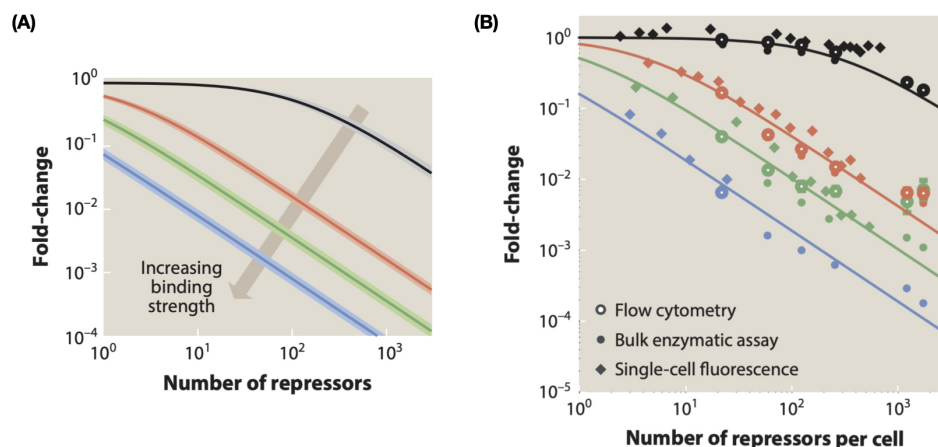


Figure 3.1: Theory meets experiment for simple repression. (A) Shows the prediction of how gene expression should change with increasing number of repressors, for four different binding sites. (B) Shows how the various data land relative to these predictions. Figure adapted from Phillips et al., 2019.

Ireland et al., 2020. For this, we are first exploring the regulation of *tetR* and *purR*, but look forward to expanding even further in the coming years. While this is still a work in progress, we are well on our way to having goals (1) and (2) addressed.

3.2 Preliminary results

Mapping barcodes

As with the Reg-Seq protocol, our experiments involve two sequencing runs: one in each we map the barcodes to the promoter regions, and one where we actually expose our cells to the experimental condition of interest and quantify the effect on gene regulation (as measured by barcode counts). Thus far, we have conducted the mapping sequencing run for both the ‘library’ containing the three wildtype operator sequences (O1, O2, and O3) as well as the O1 library mutagenized at 10%.

For the wildtype library, the coverage of barcodes is shown in Figure 3.2. As expected for a library with only three unique operator sequences, we see that we have immense coverage, with at least 8000 barcodes for each of the operators. Encouragingly, out of the 25,686 total barcodes found, just 24 (less than 0.1%) were found associated with more than one operator. While this immense number of barcodes is surely excessive, this minimal library will serve as a proof of concept and will allow us to readily assess precisely how many barcodes are needed to acquire reliable results.

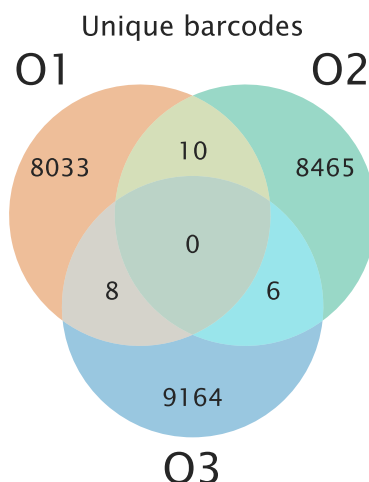


Figure 3.2: Barcode coverage for the three wildtype operator sequences. Venn diagram depicts the number of barcodes found for the O1, O2, and O3 operators. Numbers in the overlapping regions indicate barcodes that were found associated with more than one operator.

In addition to the simple wildtype library, we also mapped the the mutagenized O1 library, for which we ordered 3259 sequences from Twist Bioscience. Of these, we recovered all but 12 in the mapping run, and we even obtained an additional 619 sequences that were errors from the oligo synthesis (Figure 3.3 (A)). While errors are generally a nuisance, these single basepair mistakes actually serve to increase the diversity of our library, giving us even more sequences to work with than what we originally ordered. Furthermore, we are encourage to see that the predicted binding energies seem uniformly distributed, as we designed when selecting which mutants to order (Figure 3.3 (B)).

While we only have the first sequencing run completed for now, these preliminary results are promising with respect to having good coverage of the sequences we expected to see, and we are well-poised to analyze the data for the experimental sequencing runs as soon as they come in.

3.3 Supplementary information: library content and design

The oligo libraries used in this study were ordered from Twist Bioscience, and the complete file of sequences ordered can be found here: https://github.com/RPGroup-PBoC/Reg-Seq2/blob/master/data/twist_order/twist_order_

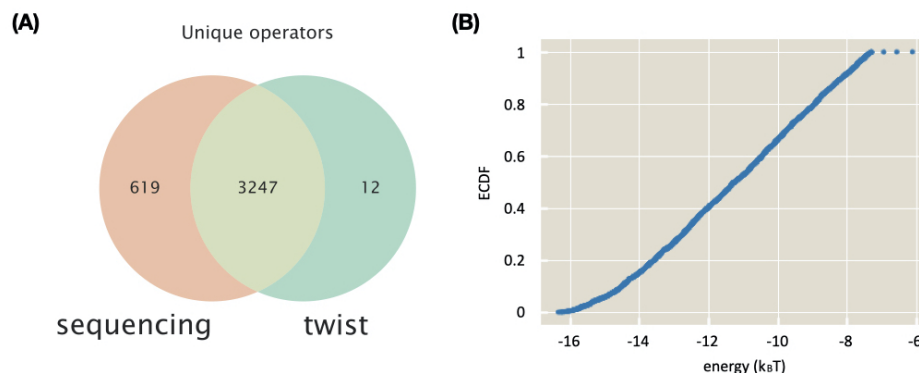


Figure 3.3: Operator coverage of the O1 library. (A) Counts for all the operators that were associated with at least 10 unique barcodes. ‘twist’ refers to those sequences which were specifically ordered, and ‘sequencing’ refers to those recovered in the mapping sequencing run. (B) Cumulative distribution function for the predicted binding energies of the operators.

full.csv. Brief descriptions of each sub-library and reference to the repository code used to create the set of sequences are as follows:

- *lacUV5*+O1 mutants at
code/experimental_design/twist_order/lacI_titration/
generate_sequences.ipynb.
- constitutive *lacUV5* mutants at
code/experimental_design/twist_order/lacUV5_mutants/
generate_sequences.ipynb.
- *lacUV5*+tetOx single and double mutants at
code/experimental_design/twist_order/tetR_regulation/
generate_sequences.ipynb.
- natural *tet* promoters mutants
code/experimental_design/twist_order/tetR_regulation/
generate_sequences.ipynb.
- *purR* simple repression single and double mutants at
code/experimental_design/twist_order/purR_titration/
purR_twist_order.ipynb.

In the library design, each sub-library was given a unique set of orthogonal primers, such that out of the total oligo pool, these individual experimental libraries could be amplified.

3.4 Supplementary information: ORBIT cloning protocol

This protocol walks through the basic steps of how to perform an ORBIT integration. This protocol is general and should work for a variety of integration plasmids and targeting oligos that can be used for many different types of insertions and deletions on the *E. coli* genome.

This protocol has been adapted from “Orbit: A new paradigm for genetic engineering of mycobacterial chromosomes” 2018. As an overview, recall that ORBIT modifications require three components that work together: the helper plasmid, an integrating plasmid, and a targeting oligo. The targeting oligo contains a 38 bp attachment sites (attB) flanked by homology arms and it gets incorporated into the genome during DNA replication directed by its homology arms. The integrating plasmid contains a 48 bp cognate attachment site (attP), which recombines with the oligo attachment site, thus integrating the entire plasmid at the location targeted in the oligo. The oligo incorporation, also called single stranded DNA recombineering (or MAGE for many oligos), is catalyzed by a single stranded DNA annealing protein (SSAP), and the modification is stabilized by temporarily suppressing the mismatch repair machinery. The recombination between the oligo attB site and the plasmid attP site is catalyzed by the site specific recombinase Bxb-1. The helper plasmid contains all of the necessary genes to achieve these reactions: SSAP - CspRecT, MutL E32K, and Bxb-1. Note that CspRecT and MutL E32K are inducibly controlled by m-toluic acid (XylS - Pm system) and Bxb-1 is separately inducible with L-arabinose (AraC - ParaB system).

Here we assume that you already have a helper plasmid strain that has been induced with m-toluic acid (CspRecT + MutL E32K) and is electrocompetent.

Materials:

1. Targeting oligo stock at 25 μ M
2. Integrating plasmid stock at \sim 100 ng/ μ L
3. Electrocompetent cells with induced helper plasmid, stored in \sim 50 μ L frozen aliquots

Protocol:

1. Consider setting up a control condition with integrating plasmid but no targeting oligo. This control will help you assess the likelihood that your ORBIT experiment worked before performing any molecular or phenotypic confirmation. It tests the background off target integration and provides a useful comparison to your + oligo conditions that should be a much higher on target integration. Typically the off target integration rate should be less than 1% of the on target integration rate.
2. Set up electroporation cuvettes and cool on ice.
3. Thaw frozen competent cells (with induced helper plasmid) on ice for 10 min.
4. Prepare the recovery culture tubes while competent cells are thawing. With sterile technique, add 30 μ L of 10% L-arabinose stock (0.1% final) and 3 mL of LB to each culture tube. You may also thaw other stock solutions during this time.

Arabinose induces the Bxb-1 integrase. Note that cells need to divide during the recovery, so 3 mL cultures allows recovery cultures to be less dense and more likely to divide than smaller volumes.

5. Add ORBIT materials to competent cells: Add \sim 100 ng of integrating plasmid to each competent cell aliquot. Add 2 μ L of targeting oligo stock (\sim 1 μ M final) to each competent cell aliquot (with the exception of the negative control).
6. Transfer to electroporation cuvettes: Gently mix competent cell mixtures and transfer to electroporation cuvettes. Tap cuvettes gently to eliminate bubbles, and replace cuvettes in ice.
7. Electroporate with typical *E. coli* settings of 1.8 kV, 25 μ F, 200 Ω . Immediately resuspend cuvette cells in 1 mL of recovery medium from the respective culture tube. Be gentle, and transfer the cells to the culture tube. Electroporate and resuspend all conditions and then proceed to recovery.
8. Recover electroporated cells: Transfer the recovery tubes to a 37°C shaker at \sim 250 rpm and recover for \sim 1 hour. Recovery time has been optimized for a 1 kb deletion with no payload on the integrating plasmid. Longer deletions and other modifications may require longer recovery periods.

9. Prepare plates: During recovery make sure agar plates are at least at room temperature. Dry plates with lids open during this time, if necessary.
10. Plate recovered cells: Plate a dilution series on kanamycin plates. Optionally, plate on LB as well as kanamycin: this allows you to calculate the overall efficiency (of all surviving cells, how many got the modification). Typically overall efficiency is ~0.5% of all cells for a 1 kb *galK* deletion.

Because of the high efficiency, 50 μ L of undiluted culture can yield a lawn of colonies, so we strongly recommend trying multiple dilutions. The preferred method is the drip plate, which can be used to accurately count 10^1 - 10^9 colonies on a single plate.

*Chapter 4***CONCLUDING THOUGHTS AND FUTURE DIRECTIONS**

In concluding, I wish to impress upon the reader the progress that has been made throughout the course of the work of my thesis, as well as future questions that this work permits the exploration of, and the limitations that still exist. In brief, I hope this work has clearly illustrated the issue of ‘regulatory ignorance’ as well as provided concrete steps toward understanding genomes.

4.1 Progress

First and foremost I would like to emphasize the progress that has been made throughout the course of my PhD, although it is important to acknowledge that none of this work was a completely solitary process. When I first joined the Phillips lab, several of the more senior graduates students in the lab were well on their way to using the Sort-Seq method to elucidate the previously unexplored regulation of a handful of genes, with their work culminating in the results shared by Belliveau et al., 2018. With these advances on the horizon, I, in collaboration with William Ireland, sought to expand the utility of Sort-Seq beyond that of a gene-by-gene endeavour.

The work done here in some ways serves as a proof of principle, but even with ‘just’ the ~100 genes explored here, we have made a non-negligible dent of a couple percent of *E. coli*’s roughly 4000 genes. This progress can in part be seen in the increase of the percent of genes with known regulation as illustrated in Figure 2.1 (34% prior to our study being completed, around 2018) as opposed to that in Figure 1.8 (38% at the time of this writing in 2021).

The work as presented here glosses over many of the difficulties we faced along the way as well as some of the intermediate steps of progress in favor of presenting the final result of over 100 genes elucidated all at once. In reality, our progress was more step-wise, first validating a library of 10 genes before tackling a full set of 100 genes. Such progress gives us hope that similar orders of magnitude of progress can be made in the coming years, allowing us to ‘finish’ the rest of the *E. coli*.

4.2 Future goals

As discussed in the introduction, the problem of regulatory ignorance extends well beyond that of *E. coli*. In fact, our regulatory knowledge drops off precipitously for higher organisms, as was illustrated in Figure 1.8. So while this specific study was done in *E. coli*, the hope is that the core principle can be extended to any number of organisms (permitted that the tools of molecular genetics are available). At first blush, there is no reason to think that the core principle of mutating regulatory regions and assessing their impact could not be carried over to other systems of interest (with some challenges, of course, as acknowledged in the following section).

In addition to expanding this work to more genes and other organisms, it's vitally important to consider precisely *what* we do with the information gained from such experiments. In this respect, I often think of the following quote from Henri Poincaré: "Science is built of facts the way a house is built of bricks: but an accumulation of facts is no more science than a pile of bricks is a house." I worry that this work may rightfully face this critique, that we are simply gathering a set of unrelated facts, without building true understanding, but I argue that there are two useful threads to be pursued upon gathering more regulatory knowledge via Reg-Seq: 1) in-depth analysis of individual genes, and 2) regulome-wide interactions.

My work so far has given a brief glimpse into what these two divergent paths may look like. On the side of exploring networks of an entire regulome, Figure 2.9 provides a small example of a single genetic circuit, where we explored both the regulation of the *arcA* gene as well as revealed the role that the ArcA protein plays in regulating another gene. While this example is quite small, it's easy to conclude that more circuits such as this one will be unveiled as more and more genes are explored in *E. coli*. So what might otherwise be seen as a disjointed set of facts may eventually produce entire genetic networks to be explored.

On the opposite side of the spectrum, the results of Reg-Seq experiment also permit the in-dept analysis of individual genes. Our very preliminary efforts to do so for both the *lac* operon as well as the *pur* and *tet* promoters was discussed in Chapter 3. While other techniques such as ChIP-Seq can permit the elucidation of where proteins are binding, they fail to give as detailed a view as the basepair-by-basepair energy matrices that result from a Reg-Seq experiment. And if we wish to not only be able to read genomes, but also write them (which can be of great interest in synthetic and bio engineered systems), this deeper understanding of *how* the protein interacts with a segment of DNA and not just that it does interact is essential.

While we have just began the process of bringing the theory to bear on a handful of genes beyond the already extensively-studied *lac* operon, I expect many more to be validated in the coming years. It is only through the careful dissection and validation of our predictions (say, of how a given mutation should impact protein binding and ultimately gene expression) that can we hope to reliably design other constructs with a desired input-output relationship at will.

4.3 Outstanding challenges

Despite the advances that have been made here, it is important to acknowledge a number of challenges that may be faced when tackling the issue of regulatory ignorance both for the rest of *E. coli* as well as for many other organisms of interest. Below I outline a handful of these present challenges, as well as some potential headway that can be made in spite of these obstacles.

Need for prior knowledge

First and foremost, it's important to acknowledge precisely how much prior knowledge went into the design of the experiments conducted here. We relied on decades of prior *E. coli* knowledge to inform precisely which regions of the genome to explore in-depth. As a concrete example, we relied on the documentation of known or predicted transcription start sites as annotated on RegulonDB (Santos-Zavaleta et al., 2019) as a guide for which regions to mutagenize. That is, we expect most regulation, especially in *E. coli*, to take place in the regions immediately surrounding the transcription start site (Rydenfelt et al., 2014a), so we prioritized these regions when designing our mutagenized sequences.

In addition to simply knowing where genes are being transcribed from, we also relied on existing proteomic data (such as from Schmidt et al., 2016) to get a hint at both which genes may be regulated and also which environmental growth conditions influence their expression. Such data was presented in Figure 1.9, and almost taken for granted, but this information was essential to unearthing the regulation at play, and comparable data may not yet exist for other organisms.

Sequences and sequencing limitations

As alluded to in the previous section, in the current iteration of Reg-Seq, we had to be fairly decisive with which regions to explore with our library mutagenized promoter segments. Such targeted mutagenesis allowed us to get at the heart of the regulation of a given gene without having to approach the genes blindly. As a

small back-of-the-envelope estimate, if we were to explore the entire 4.6 Mb *E.coli* genome with our fairly small 160 bp regions, we would have to prescribe nearly 30,000 promoter regions. In other words, such an endeavour would require the work done here (on around 100 such 160 bp regions) to be done 300 times over, which might be unattainable at least with the current technologies. However, as sequencing approaches continue to improve, I suspect this to become less of a bottleneck.

Even with the current constraints on sequencing (financial or otherwise), work from others provides a way forward that does not rely on blindly examining all regions of the genome at this perhaps excessive basepair-by-pair resolution. In a related approach, Urtecho et al., 2020 were able to perturb binding sites wholesale by mutating regions of ten consecutive basepairs (as opposed to our scattered single-basepair mutations). With this more blunt approach, they were able to tile these mutated 10 bp chunks across all the entire genome. While such an approach cannot yield a predictive energy matrix or other quantitative measures of how a given protein binds to DNA, it is the perfect discovery tool for finding precisely where regulation seems to be occurring. Moving forward, I envision using an approach akin to Urtecho et al., 2020, followed by a more in-depth Reg-Seq approach upon the regions that were discovered by the broader method.

What are the relevant environmental conditions?

Perhaps the largest issue when it comes to discovering gene regulation writ large is knowing which environmental condition (or conditions) are most relevant to a given gene's expression. That is, just because we fail to capture gene regulation even across a wide range of environmental conditions, that does not imply that a given gene is necessarily *not* regulated, but that we may have failed to test in a relevant condition such that the regulation would be enacted.

This issue is concisely posed by the title of work by Lindsley and Rutter, 2006: "Whence cometh the allosterome?". That is, the modern era of molecular biology is filled with various '-omes': the genome, the transcriptome, the proteome, etc. But even with the development of all these sequencing approaches, until recently, it seemed out of reach that we would be able to readily detect all the possible allosteric effectors (i.e. ligands) of a given protein. However, progress has been made even here, as shown by the work of Piazza et al., 2018, where they are able to detect which protein regions are bound to metabolites, through an approach akin to footprinting used to detect protein-DNA interactions. By cross-linking proteins under different

environmental conditions (e.g. in the presence of different metabolites), exposing the proteins to protease digestion, and then quantifying the digested samples via mass spectrometry, they were able to detect differential abundance of protein regions. That is, a region of a protein bound by metabolite will be less amenable to protease, causing the corresponding amino acid sequence in that region of the protein to be less abundant in the final quantification. While Piazza et al., 2018 specifically looked at the role of metabolites, I can imagine this being expanded to a wider range of potential effectors, including those that interact with transcription factors.

Challenges in other organisms

The issues mentioned above become more pertinent when considering organisms beyond *E. coli*. That is, as we move to less well-studied organisms, there will inherently be less information to build off. Thankfully, what has taken decades to be elucidated in *E. coli* can hopefully be achieved in much faster time scales for other organisms. The recent development of various ‘atlases’, such as The Human Cell Atlas, The Human Protein Atlas, The Tabula Muris Consortium, and others, suggest that such foundational information may very well become more readily available in the coming years. Even for a completely new organism, the prospect of getting the genome, transcriptome, and proteome within a month’s time is becoming evermore reasonable.

In addition to having less current information to go off of for other organisms, it’s also the case that regulation is simply much more complicated in higher organisms. With enhancers often enacting their regulation from kilobases, and sometimes even megabases away, it is clear that our approach of selecting single 160 bp regions will not suffice in many eukaryotic systems. This problem of regulation occurring at a distance likely poses the largest hurdle in developing Reg-Seq in other organisms. However, I remain optimistic that advances in technology will trivialize this problem in the coming decades, if not years. Even with just existing technologies, I imagine it could be possible to pair Reg-Seq on a proximal regulatory region with some sort of chromatin conformation capture technology to assess the distal regulatory regions simultaneously. And we might just be able to piece together how specific transcription factors bind to enhancer regions, with the same level of detail as we have here for promoters in *E. coli*.

In closing, despite the caveats discussed here, I feel that many of the challenges posed here present solvable problems, especially as sequencing approaches and technologies advance in the coming years. With this perspective, my cautiously optimistic hope is that this work and others like it have set a foundation such that we might be able to truly say that we as a field “understand genomes” at some point in the coming decades of my scientific career.

BIBLIOGRAPHY

- Ackers, G. K., A. D. Johnson, and M. A. Shea (1982). “Quantitative model for gene regulation by lambda phage repressor”. en. In: *Proceedings of the National Academy of Sciences of the United States of America*, p. 5. DOI: 10.1073/pnas.79.4.1129.
- Arsène, F., T. Tomoyasu, and B. Bukau (Apr. 2000). “The heat shock response of *Escherichia coli*”. en. In: *International Journal of Food Microbiology* 55.1-3, pp. 3–9. ISSN: 01681605. DOI: 10.1016/S0168-1605(00)00206-3. (Visited on 06/19/2020).
- Barnes, S. L., N. M. Belliveau, W. T. Ireland, J. B. Kinney, and R. Phillips (2019). “Mapping DNA sequence to transcription factor binding energy *in vivo*”. In: *PLoS Computational Biology* 15.2, pp. 1–29. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1006226.
- Belliveau, N. M., S. L. Barnes, W. T. Ireland, D. L. Jones, M. J. Sweredoski, A. Moradian, S. Hess, J. B. Kinney, and R. Phillips (2018). “Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria”. In: *Proceedings of the National Academy of Sciences of the United States of America* 115.21, E4796–E4805. DOI: 10.1073/pnas.1722055115.
- Benjamini, Y. and Y. Hochberg (Jan. 1995). “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. en. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1, pp. 289–300. ISSN: 00359246. DOI: 10.1111/j.2517-6161.1995.tb02031.x.
- Berg, O. G. and P. H. von Hippel (1987). “Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters”. In: *Journal of Molecular Biology* 193.4, pp. 723–50. DOI: 10.1016/0022-2836(87)90354-8.
- Bintu, L., N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips (Apr. 2005). “Transcriptional regulation by the numbers: models”. en. In: *Current Opinion in Genetics & Development*. Chromosomes and expression mechanisms 15.2, pp. 116–124. DOI: 10.1016/j.gde.2005.02.007.
- Blattner, F. R. (Sept. 1997). “The Complete Genome Sequence of *Escherichia coli* K-12”. en. In: *Science* 277.5331, pp. 1453–1462. ISSN: 00368075, 10959203. DOI: 10.1126/science.277.5331.1453. (Visited on 01/06/2020).
- Bremer, H. and P. Dennis (1996). “Modulation of Chemical Composition and Other Parameters of the Cell by Growth Rate”. In: *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*. Ed. by F. C. Neidhardt. 2nd ed. Chap. 97.

- Brewster, R. C., D. L. Jones, and R. Phillips (Dec. 2012). “Tuning Promoter Strength through RNA Polymerase Binding Site Design in *Escherichia coli*”. en. In: *PLoS Computational Biology* 8.12. Ed. by E. van Nimwegen, e1002811. DOI: 10.1371/journal.pcbi.1002811.
- Brewster, R. C., F. M. Weinert, H. G. Garcia, D. Song, M. Rydenfelt, and R. Phillips (Mar. 2014). “The Transcription Factor Titration Effect Dictates Level of Gene Expression”. en. In: *Cell* 156.6, pp. 1312–1323. ISSN: 0092-8674. DOI: 10.1016/j.cell.2014.02.022. URL: <http://www.sciencedirect.com/science/article/pii/S0092867414002219> (visited on 06/06/2020).
- Browning, D. F. and S. J. W. Busby (2016). “Local and global regulation of transcription initiation in bacteria”. In: *Nature Reviews Microbiology*, pp. 638–650. ISSN: 1740-1526. DOI: 10.1038/nrmicro.2016.103.
- Buchler, N. E., U. Gerland, and T. Hwa (Apr. 2003). “On schemes of combinatorial transcription logic”. In: *Proceedings of the National Academy of Sciences of the United States of America* 100.9, pp. 5136–5141. DOI: 10.1073/pnas.0930314100.
- Chure, G., M. Razo-Mejia, N. M. Belliveau, T. Einav, Z. A. Kaczmarek, S. L. Barnes, M. Lewis, and R. Phillips (2019). “Predictive shifts in free energy couple mutations to their phenotypic consequences”. In: *Proceedings of the National Academy of Sciences of the United States of America* 116.37, pp. 18275–18284. DOI: 10.1073/pnas.1907869116.
- Compan, I. and D. Touati (1994). “Anaerobic activation of *arcA* transcription in *Escherichia coli*: roles of Fnr and ArcA”. en. In: *Molecular Microbiology* 11.5, pp. 955–964. ISSN: 1365-2958. DOI: 10.1111/j.1365-2958.1994.tb00374.x.
- Conway, T., J. P. Creecy, S. M. Maddox, J. E. Grissom, T. L. Conkle, T. M. Shadid, J. Teramoto, P. San Miguel, T. Shimada, A. Ishihama, H. Mori, and B. L. Wanner (July 2014). “Unprecedented High-Resolution View of Bacterial Operon Architecture Revealed by RNA Sequencing”. en. In: *mBio* 5.4. Ed. by S. Adhya, e01442–14. ISSN: 2150-7511. DOI: 10.1128/mBio.01442-14. (Visited on 11/08/2019).
- Cox, J. and M. Mann (Dec. 2008). “MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification”. en. In: *Nature Biotechnology* 26.12, pp. 1367–1372. ISSN: 1546-1696. DOI: 10.1038/nbt.1511. (Visited on 01/18/2020).
- Cox, J., I. Matic, M. Hilger, N. Nagaraj, M. Selbach, J. V. Olsen, and M. Mann (May 2009). “A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics”. en. In: *Nature Protocols* 4.5, pp. 698–705. ISSN: 1754-2189, 1750-2799. DOI: 10.1038/nprot.2009.36. (Visited on 01/06/2020).

- Datsenko, K. A. and B. L. Wanner (2000). “One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products”. In: *Proceedings of the National Academy of Sciences of the United States of America* 97.12, pp. 6640–6645. ISSN: 00278424. DOI: 10.1073/pnas.120163297.
- Forcier, T. L., A. Ayaz, M. S. Gill, D. Jones, R. Phillips, and J. B. Kinney (2018). “Measuring cis-regulatory energetics in living cells using allelic manifolds”. In: *eLife*. ISSN: 2050084X. DOI: 10.7554/eLife.40618.
- Fulco, C. P., J. Nasser, T. R. Jones, G. Munson, D. T. Bergman, V. Subramanian, S. R. Grossman, R. Anyoha, B. R. Doughty, T. A. Patwardhan, T. H. Nguyen, M. Kane, E. M. Perez, N. C. Durand, C. A. Lareau, E. K. Stamenova, E. L. Aiden, E. S. Lander, and J. M. Engreitz (2019). “Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations”. In: *Nature Genetics* 51.12, pp. 1664–1669. DOI: 10.1038/s41588-019-0538-0.
- Galstyan, V., L. Funk, T. Einav, and R. Phillips (2019). “Combinatorial Control through Allostery”. In: *The Journal of Physical Chemistry B* 123, pp. 2792–2800. DOI: 10.1021/acs.jpcc.8b12517.
- Gama-Castro, S. et al. (2016). “RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond”. In: *Nucleic Acids Research* 44.D1, pp. D133–D143. DOI: 10.1093/nar/gkv1156.
- Gao, Y., J. T. Yurkovich, S. W. Seo, I. Kabimoldayev, K. Chen, A. V. Sastry, X. Fang, N. Mih, L. Yang, J. Eichner, B.-k. Cho, D. Kim, and B. O. Palsson (2018). “Systematic discovery of uncharacterized transcription factors in *Escherichia coli* K-12 MG1655”. In: *Nucleic Acids Research* 46.20, pp. 10682–10696. DOI: 10.1093/nar/gky752.
- Garcia, H. G. and R. Phillips (July 2011). “Quantitative dissection of the simple repression input-output function”. en. In: *Proceedings of the National Academy of Sciences of the United States of America* 108.29, pp. 12173–12178. DOI: 10.1073/pnas.1015616108. (Visited on 12/08/2019).
- Ghatak, S., Z. A. King, A. Sastry, and B. O. Palsson (2019). “The y-ome defines the 35% of *Escherichia coli* genes that lack experimental evidence of function”. In: *Nucleic Acids Research* 47.5, pp. 2446–2454. ISSN: 13624962. DOI: 10.1093/nar/gkz030.
- Goodall, E. C. A., A. Robinson, I. G. Johnston, S. Jabbari, K. A. Turner, A. F. Cunningham, P. A. Lund, J. A. Cole, and I. R. Henderson (2018). “The Essential Genome of *Escherichia coli* K-12”. In: *mBio* 9.1. DOI: 10.1128/mBio.02096-17.
- Goodman, J. and J. Weare (Jan. 2010). “Ensemble samplers with affine invariance”. en. In: *Communications in Applied Mathematics and Computational Science* 5.1, pp. 65–80. ISSN: 2157-5452, 1559-3940. DOI: 10.2140/camcos.2010.5.65. (Visited on 06/01/2020).

- Grass, G., S. Franke, N. Taudte, D. H. Nies, L. M. Kucharski, M. E. Maguire, and C. Rensing (Mar. 2005). “The Metal Permease ZupT from *Escherichia coli* Is a Transporter with a Broad Substrate Spectrum”. In: *Journal of Bacteriology* 187.5, pp. 1604–1611. DOI: 10.1128/JB.187.5.1604-1611.2005.
- Gupta, S., J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble (2007). “Quantifying similarity between motifs”. In: *Genome Biology* 8.2. DOI: 10.1186/gb-2007-8-2-r24.
- Han, L., H. G. Garcia, S. Blumberg, K. B. Towles, J. F. Beausang, P. C. Nelson, and R. Phillips (2009). “Concentration and Length Dependence of DNA Looping in Transcriptional Regulation”. In: *PLoS ONE* 4.5. ISSN: 19326203. DOI: 10.1371/journal.pone.0005621. arXiv: 0806.1860.
- Hannon, G. J. (2010). *FASTX-Toolkit*. URL: http://hannonlab.cshl.edu/fastx_toolkit/ (visited on 06/20/2020).
- Hastie, T., R. Tibshirani, and J. Friedman (Aug. 2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. en. 2nd ed. Springer Science & Business Media. ISBN: 978-0-387-84858-7.
- Hirokawa, S., G. Chure, N. M. Belliveau, G. A. Lovely, M. Anaya, D. G. Schatz, D. Baltimore, and R. Phillips (2020). “Sequence-dependent dynamics of synthetic and endogenous RSSs in V(D)J recombination”. In: *Nucleic Acids Research* D, pp. 1–14. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa418.
- Huang, M., K. S., X. Lu, S. Lu, Q. Shen, R. Wang, J. Gao, Y. Hong, Q. Li, D. Ni, J. Xu, G. Chen, and J. Zhang (2018). “AlloFinder: a strategy for allosteric modulator discovery and allosterome analyses”. In: *Nucleic Acids Research* 46.W1, W451–W458.
- Huerta, A. M. and J. Collado-Vides (Oct. 2003). “Sigma70 Promoters in *Escherichia coli*: Specific Transcription in Dense Regions of Overlapping Promoter-like Signals”. en. In: *Journal of Molecular Biology* 333.2, pp. 261–278. DOI: 10.1016/j.jmb.2003.07.017. (Visited on 06/18/2020).
- Ireland, W. T., S. M. Beeler, E. Flores-Bautista, N. S. McCarty, T. Röschinger, N. M. Belliveau, M. J. Sweredoski, A. Moradian, J. B. Kinney, and R. Phillips (2020). “Deciphering the regulatory genome of *Escherichia coli*, one hundred promoters at a time”. In: *eLife*. ISSN: 2050084X. DOI: 10.7554/ELIFE.55308. arXiv: 2001.07396.
- Ireland, W. T. and J. B. Kinney (May 2016). “MPAthic: Quantitative Modeling of Sequence-Function Relationships for massively parallel assays”. en. In: *bioRxiv*. DOI: 10.1101/054676. (Visited on 01/15/2020).
- Jacob, F. and J. Monod (1961). “On the Regulation of Gene Activity”. en. In: *Cold Spring Harbor Symposia on Quantitative Biology* 26, p. 19. DOI: 10.1101/SQB.1961.026.01.024.

- Johnson, C. M. and R. F. Schleif (June 1995). “In vivo induction kinetics of the arabinose promoters in *Escherichia coli*.” In: *Journal of Bacteriology* 177.12, pp. 3438–3442. ISSN: 0021-9193. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC177046/> (visited on 07/20/2020).
- Johnson, S., M. Lindén, and R. Phillips (2012). “Sequence dependence of transcription factor-mediated DNA looping”. In: *Nucleic Acids Research* 40.16, pp. 7728–7738. ISSN: 03051048. DOI: 10.1093/nar/gks473.
- Kargeti, M. and K. V. Venkatesh (2017). “The effect of global transcriptional regulators on the anaerobic fermentative metabolism of *Escherichia coli*”. en. In: *Molecular BioSystems* 13.7, pp. 1388–1398. ISSN: 1742-206X, 1742-2051. DOI: 10.1039/C6MB00721J. (Visited on 01/16/2020).
- Keseler, I. M. et al. (2017). “The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12”. In: *Nucleic Acids Research* 45.D1, pp. D543–D550. ISSN: 0305-1048. DOI: 10.1093/nar/gkw1003.
- Kinney, J. B. and G. S. Atwal (Jan. 2014). “Parametric Inference in the Large Data Limit Using Maximally Informative Models”. In: *Neural Computation* 26.4, pp. 637–653. ISSN: 0899-7667. DOI: 10.1162/NECO_a_00568. (Visited on 06/06/2020).
- Kinney, J. B. and D. M. McCandlish (2019). “Massively Parallel Assays and Quantitative Sequence-Function Relationships”. In: *Annual Review of Genomics and Human Genetics* 20.1, pp. 99–127. DOI: 10.1146/annurev-genom-083118-014845.
- Kinney, J. B., A. Murugan, C. G. Callan, and E. C. Cox (2010). “Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence”. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.20, pp. 9158–9163. DOI: 10.1073/pnas.1004290107.
- Körner, H., H. J. Sofia, and W. G. Zumft (Dec. 2003). “Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs”. en. In: *FEMS Microbiology Reviews* 27.5, pp. 559–592. ISSN: 1574-6976. DOI: 10.1016/S0168-6445(03)00066-4. (Visited on 01/16/2020).
- Kosuri, S., D. B. Goodman, G. Cambray, V. K. Mutalik, and Y. Gao (2013). “Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*”. In: *Proceedings of the National Academy of Sciences of the United States of America* 110.34. DOI: 10.1073/pnas.1301301110.
- Kwasnieski, J. C., I. Mogno, C. A. Myers, J. C. Corbo, and B. A. Cohen (2012). “Complex effects of nucleotide variants in a mammalian cis-regulatory element”. In: *Proceedings of the National Academy of Sciences of the United States of America* 109.47, pp. 19498–503. DOI: 10.1073/pnas.1210678109.

- Larson, T. J., J. S. Cantwell, and A. T. van Loo-Bhattacharya (Mar. 1992). “Interaction at a Distance Between Multiple Operators Controls the Adjacent, Divergently Transcribed glpTQ-glpACB Operons of *Escherichia coli* K-12”. eng. In: *The Journal of Biological Chemistry* 267.9, pp. 6114–6121. ISSN: 0021-9258.
- Larson, T. J., S. Ye, D. L. Weissenborn, and H. J. Hoffmann (1987). “Purification and Characterization of the Repressor for the *sn*-Glycerol 3-Phosphate Regulon of *Escherichia coli* K12”. In: *Journal of Biological Chemistry* 262.33, pp. 15869–15874.
- Li, G.-Y., Y. Zhang, M. Inouye, and M. Ikura (June 2008). “Structural mechanism of transcriptional autorepression of the *Escherichia coli* RelB/RelE antitoxin/toxin module”. eng. In: *Journal of Molecular Biology* 380.1, pp. 107–119. ISSN: 1089-8638. DOI: 10.1016/j.jmb.2008.04.039.
- Lin, E. C. C. (1976). “Glycerol Dissimilation and Its Regulation in Bacteria”. In: *Annual Review of Microbiology* 30.1, pp. 535–578. DOI: 10.1146/annurev.mi.30.100176.002535.
- Lindsley, J. E. and J. Rutter (2006). “Whence cometh the allosterome?” In: *Proceedings of the National Academy of Sciences of the United States of America* 103.28, pp. 10533–10535. ISSN: 00278424. DOI: 10.1073/pnas.0604452103.
- Lister, R., R. C. O’Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, and J. R. Ecker (2008). “Highly Integrated Single-Base Resolution Maps of the Epigenome in *Arabidopsis*”. In: *Cell*. ISSN: 00928674. DOI: 10.1016/j.cell.2008.03.029.
- Lovely, G. A., R. C. Brewster, D. G. Schatz, D. Baltimore, and R. Phillips (2015). “Single-molecule analysis of RAG-mediated V(D)J DNA cleavage”. In: *Proceedings of the National Academy of Sciences of the United States of America* 112.14. Publisher: National Academy of Sciences, E1715–E1723. ISSN: 0027-8424. (Visited on 07/19/2020).
- Lutkenhaus, J. (2007). “Assembly Dynamics of the Bacterial MinCDE System and Spatial Regulation of the Z Ring”. In: *Annual Review of Biochemistry* 76.1, pp. 539–562. DOI: 10.1146/annurev.biochem.75.103004.142652.
- Lutz, R. and H. Bujard (1997). “Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I₁-I₂ regulatory elements”. In: *Nucleic Acids Research* 25.6, pp. 1203–1210. DOI: 10.1093/nar/25.6.1203.
- Magoč, T. and S. L. Salzberg (Nov. 2011). “FLASH: fast length adjustment of short reads to improve genome assemblies”. en. In: *Bioinformatics* 27.21, pp. 2957–2963. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr507. (Visited on 01/06/2020).

- Melnikov, A., A. Murugan, X. Zhang, T. Tesileanu, L. Wang, P. Rogov, S. Feizi, A. Gnirke, C. G. C. Jr, J. B. Kinney, M. Kellis, E. S. Lander, and T. S. Mikkelsen (2012). “Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay”. In: *Nature Biotechnology* 30.3, pp. 271–277. ISSN: 1087-0156. DOI: 10.1038/nbt.2137.
- Mendoza-Vargas, A., L. Olvera, M. Olvera, R. Grande, L. Vega-Alvarado, B. Taboada, V. Jimenez-Jacinto, H. Salgado, K. Juárez, B. Contreras-Moreira, A. M. Huerta, J. Collado-Vides, and E. Morett (Oct. 2009). “Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in *E. coli*”. en. In: *PLoS ONE* 4.10. Ed. by C. Creighton, e7526. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0007526. (Visited on 01/11/2020).
- Mittler, G., F. Butter, and M. Mann (2009). “A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements”. In: *Genome Research* 19.2, pp. 284–93. DOI: 10.1101/gr.081711.108.
- Monod, J. (1966). “From enzymatic adaptation to allosteric transitions”. In: *Science* 154.3748, pp. 475–483. ISSN: 00368075. DOI: 10.1126/science.154.3748.475.
- Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold (2008). “Mapping and quantifying mammalian transcriptomes by RNA-Seq”. In: *Nature Methods* 5.7, pp. 621–628. ISSN: 1548-7105. DOI: 10.1038/nmeth.1226.
- Myers, K. S., H. Yan, I. M. Ong, D. Chung, K. Liang, F. Tran, S. Keleş, R. Landick, and P. J. Kiley (June 2013). “Genome-scale Analysis of *Escherichia coli* FNR Reveals Complex Features of Transcription Factor Binding”. In: *PLOS Genetics* 9.6, pp. 1–24. DOI: 10.1371/journal.pgen.1003565.
- Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder (2008). “The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing”. In: *Science*. ISSN: 00368075. DOI: 10.1126/science.1158441.
- “Orbit: A new paradigm for genetic engineering of mycobacterial chromosomes” (2018). In: *mBio* 9.6, pp. 1–20. ISSN: 21507511.
- Pappireddi, N., L. Martin, and M. Wühr (2019). “A Review on Quantitative Multiplexed Proteomics”. In: *ChemBioChem* 20.10, pp. 1210–1224. ISSN: 14397633. DOI: 10.1002/cbic.201800650.
- Partridge, J. D., D. M. Bodenmiller, M. S. Humphrys, and S. Spiro (2009). “NsrR targets in the *Escherichia coli* genome: new insights into DNA sequence requirements for binding and a role for NsrR in the regulation of motility”. en. In: *Molecular Microbiology* 73.4, pp. 680–694. ISSN: 1365-2958. DOI: 10.1111/j.1365-2958.2009.06799.x. (Visited on 01/16/2020).
- Patil, A., D. Huard, and C. J. Fonnesbeck (July 2010). “PyMC: Bayesian Stochastic Modelling in Python”. In: *Journal of Statistical Software* 35.4, pp. 1–81. ISSN: 1548-7660. (Visited on 06/05/2020).

- Patwardhan, R. P., J. B. Hiatt, D. M. Witten, M. J. Kim, R. P. Smith, D. May, C. Lee, J. M. Andrie, S. I. Lee, G. M. Cooper, N. Ahituv, L. A. Pennacchio, and J. Shendure (2012). “Massively parallel functional dissection of mammalian enhancers *in vivo*”. In: *Nature Biotechnology* 30.3, pp. 265–70. doi: 10.1038/nbt.2136.
- Patwardhan, R. P., C. Lee, O. Litvin, D. L. Young, D. Pe’er, and J. Shendure (2009). “High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis”. In: *Nature Biotechnology* 27.12, pp. 1173–1175. doi: 10.1038/nbt.1589.
- Phillips, R., N. M. Belliveau, G. Chure, H. G. Garcia, M. Razo-Mejia, and C. Scholes (2019). “Figure 1 Theory Meets Figure 2 Experiments in the Study of Gene Expression”. In: *Annual Review of Biophysics* 48, pp. 121–163. doi: 10.1146/annurev-biophys-052118-115525.
- Piazza, I., K. Kochanowski, V. Cappelletti, T. Fuhrer, E. Noor, U. Sauer, and P. Picotti (2018). “A Map of Protein-Metabolite Interactions Reveals Principles of Chemical Communication”. In: *Cell* 172.1-2, pp. 358–372. doi: 10.1016/j.cell.2017.12.006.
- Razo-Mejia, M., S. L. Barnes, N. M. Belliveau, G. Chure, T. Einav, M. Lewis, and R. Phillips (2018). “Tuning Transcriptional Regulation through Signaling: A Predictive Theory of Allosteric Induction”. In: *Cell Systems* 6.4, 456–469.e10. ISSN: 24054720. doi: 10.1016/j.cels.2018.02.004. arXiv: 1702.07460.
- Rhee, K. Y., D. F. Senear, and G. W. Hatfield (May 1998). “Activation of Gene Expression by a Ligand-induced Conformational Change of a Protein-DNA Complex”. en. In: *Journal of Biological Chemistry* 273.18, pp. 11257–11266. ISSN: 0021-9258, 1083-351X. doi: 10.1074/jbc.273.18.11257.
- Rydenfelt, M., H. G. Garcia, R. S. Cox III, and R. Phillips (2014a). “The Influence of Promoter Architectures and Regulatory Motifs on Gene Expression in *Escherichia coli*”. In: *PLoS One* 9.12, e114347. doi: 10.1371/journal.pone.0121935.
- Rydenfelt, M., R. S. Cox III, H. Garcia, and R. Phillips (Jan. 2014b). “Statistical mechanical model of coupled transcription from multiple promoters due to transcription factor titration”. en. In: *Physical Review E* 89.1, p. 012702. doi: 10.1103/PhysRevE.89.012702. (Visited on 06/06/2020).
- Santos-Zavaleta, A., H. Salgado, S. Gama-Castro, M. Sánchez-Pérez, L. Gómez-Romero, D. Ledezma-Tejeda, J. S. García-Sotelo, K. Alquicira-Hernández, L. J. Muñoz-Rascado, P. Peña-Loredo, C. Ishida-Gutiérrez, D. A. Velázquez-Ramírez, V. D. Moral-Chávez, C. Bonavides-Martínez, C.-F. Méndez-Cruz, J. Galagan, and J. Collado-Vides (2019). “RegulonDB v 10.5: Tackling Challenges to Unify Classic and High Throughput Knowledge of Gene Regulation in *E. coli* K-12”. In: *Nucleic Acids Research* 47, pp. 212–220. doi: 10.1093/nar/gky1077.

- Schmidt, A., K. Kochanowski, S. Vedelaar, E. Ahrné, B. Volkmer, L. Callipo, K. Knoop, M. Bauer, R. Aebersold, and M. Heinemann (2016). “The quantitative and condition-dependent *Escherichia coli* proteome”. In: *Nature Biotechnology* 34.1, pp. 104–110. doi: 10.1038/nbt.3418.
- Schneider, T. D. and R. Stephens (1990). “Sequence logos: a new way to display consensus sequences”. en. In: *Nucleic Acids Research* 18.20, pp. 6097–6100. ISSN: 0305-1048, 1362-4962. doi: 10.1093/nar/18.20.6097. (Visited on 12/14/2019).
- Schweizer, H., W. Boos, and T. J. Larson (1985). “Repressor for the *sn*-glycerol-3-phosphate regulon of *Escherichia coli* K-12: cloning of the *glpR* gene and identification of its product”. en. In: *Journal of Bacteriology* 161.2, pp. 563–566. ISSN: 0021-9193, 1098-5530. doi: 10.1128/JB.161.2.563-566.1985. (Visited on 01/16/2020).
- Seoh, H. K. and P. C. Tai (Mar. 1999). “Catabolic repression of *secB* expression is positively controlled by cyclic AMP (cAMP) receptor protein-cAMP complexes at the transcriptional level”. eng. In: *Journal of Bacteriology* 181.6, pp. 1892–1899. ISSN: 0021-9193.
- Sharon, E., Y. Kalma, A. Sharp, T. Raveh-Sadka, M. Levo, D. Zeevi, L. Keren, Z. Yakhini, A. Weinberger, and E. Segal (2012). “Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters”. In: *Nature Biotechnology* 30.6, pp. 521–30. doi: 10.1038/nbt.2205.
- Skene, P. J. and S. Henikoff (June 2015). “A simple method for generating high-resolution maps of genome-wide protein binding”. In: *eLife* 4, e09225. doi: 10.7554/eLife.09225.
- Stormo, G. D. and D. S. Fields (1998). “Specificity, free energy and information content in protein-DNA interactions”. In: *Trends in Biochemical Sciences* 23.3, pp. 109–13. doi: 10.1016/S0968-0004(98)01187-6.
- Stuart, T. and R. Satija (2019). “Integrative single-cell analysis”. In: *Nature Reviews Genetics* 20, pp. 257–272. ISSN: 1471-0064. doi: 10.1038/s41576-019-0093-7.
- Tareen, A. and J. B. Kinney (2019). “Biophysical models of cis-regulation as interpretable neural networks”. In: *bioRxiv*. doi: 10.1101/835942.
- Urtecho, G., K. Insigne, A. D. Tripp, M. Brinck, N. B. Lubock, H. Kim, T. Chan, and S. Kosuri (2020). “Genome-wide Functional Characterization of *Escherichia coli* Promoters and Regulatory Elements Responsible for their Function”. In: *bioRxiv*. doi: 10.1101/2020.01.04.894907.
- Urtecho, G., A. D. Tripp, K. D. Insigne, H. Kim, and S. Kosuri (2019). “Systematic Dissection of Sequence Elements Controlling σ 70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in *Escherichia coli*”. In: *Biochemistry* 58.11, pp. 1539–1551. doi: 10.1021/acs.biochem.7b01069.

- Vilar, J. M., C. C. Guet, and S. Leibler (2003). “Modeling network dynamics: the *lac* operon, a case study”. In: *Journal of Cell Biology* 161.3, pp. 471–6. DOI: 10.1083/jcb.200301125.
- Vilar, J. M. and S. Leibler (2003). “DNA Looping and Physical Constraints on Transcription Regulation”. In: *Journal of Molecular Biology* 331.5, pp. 981–9. DOI: 10.1016/S0022-2836(03)00764-2.
- Vilar, J. M. and L. Saiz (2013). “Reliable Prediction of Complex Phenotypes from a Modular Design in Free Energy Space: An Extensive Exploration of the *lac* Operon”. In: *ACS Synthetic Biology* 2.10, pp. 576–86. DOI: 10.1021/sb400013w.
- Weinert, F. M., R. C. Brewster, M. Rydenfelt, R. Phillips, and W. K. Kegel (Dec. 2014). “Scaling of Gene Expression with Transcription-Factor Fugacity”. In: *Physical Review Letters* 113.25, p. 258101. DOI: 10.1103/PhysRevLett.113.258101. (Visited on 06/06/2020).
- Weissenborn, D. L., N. Wittekindt, and T. J. Larson (Mar. 1992). “Structure and Regulation of the *glpFK* Operon Encoding Glycerol Diffusion Facilitator and Glycerol Kinase of *Escherichia coli* K-12”. eng. In: *The Journal of Biological Chemistry* 267.9, pp. 6122–6131. ISSN: 0021-9258.
- Yamamoto, N., K. Nakahigashi, T. Nakamichi, M. Yoshino, Y. Takai, Y. Touda, A. Furubayashi, S. Kinjo, H. Dose, M. Hasegawa, K. A. Datsenko, T. Nakayashiki, M. Tomita, B. L. Wanner, and H. Mori (2009). “Update on the Keio collection of *Escherichia coli* single-gene deletion mutants”. In: *Molecular Systems Biology* 5, pp. 335–335. DOI: 10.1038/msb.2009.92.
- Yang, B. and T. J. Larson (Dec. 1996). “Action at a distance for negative control of transcription of the *glpD* gene encoding *sn*-glycerol 3-phosphate dehydrogenase of *Escherichia coli* K-12”. eng. In: *Journal of Bacteriology* 178.24, pp. 7090–7098. ISSN: 0021-9193. DOI: 10.1128/jb.178.24.7090-7098.1996.
- Ye, S. Z. and T. J. Larson (Sept. 1988). “Structures of the promoter and operator of the *glpD* gene encoding aerobic *sn*-glycerol-3-phosphate dehydrogenase of *Escherichia coli* K-12”. eng. In: *Journal of Bacteriology* 170.9, pp. 4209–4215. ISSN: 0021-9193. DOI: 10.1128/jb.170.9.4209-4215.1988.
- Zaslaver, A., A. Bren, M. Ronen, S. Itzkovitz, I. Kikoin, S. Shavit, W. Liebermeister, M. Surette, and U. Alon (2006). “A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*”. In: *Nature Methods* 8.3, pp. 623–628. DOI: 10.1038/nmeth895.
- Zhao, N., W. Oh, D. Trybul, K. S. Thrasher, T. J. Kingsbury, and T. J. Larson (Apr. 1994). “Characterization of the interaction of the *glp* repressor of *Escherichia coli* K-12 with single and tandem *glp* operator variants”. eng. In: *Journal of Bacteriology* 176.8, pp. 2393–2397. ISSN: 0021-9193. DOI: 10.1128/jb.176.8.2393-2397.1994.